Etude comportementale des mesures d'intérêt d'extraction de connaissances

Dhouha Grissa

N° D'ORDRE: D. U. 2401 EDSPIC: 626

Université Blaise Pascal
U.F.R. Sciences et Technologies de l'information
École Doctorale des Sciences de l'Ingénieur

Université Tunis-El Manar
Faculté des Sciences de Tunis, Département Informatique
École Doctorale Mathématiques, Informatique, Sciences et Technologie de la Matière

T H È S E

En cotutelle

Présentée par

Dhouha GRISSA

soutenue publiquement à Clermont-Ferrand le 2 décembre 2013

en vue de l'obtention du

Doctorat en Informatique

Étude comportementale des mesures d'intérêt d'extraction de connaissances

MEMBRES du JURY :

Mme. Amel BORGI	Maitre de conférences, HDR	Université Tunis-El-Manar	(***Rapporteur***)
M. Jean DIATTA	Professeur	Université de la Réunion	(***Rapporteur***)
M. Richard EMILION	Professeur	Université d'Orléans	(***Examinateur***)
M. Mohamed Mohsen GAMMOUDI	Professeur	Université de la Manouba	(***Examinateur***)
M. Sadok BEN YAHIA	Professeur	Université Tunis-El-Manar	(***Directeur de thèse***)
M. Engelbert MEPHU NGUIFO	Professeur	Université Blaise-Pascal	(***Directeur de thèse***)
Mme. Sylvie GUILLAUME	Maitre de conférences	Université d'Auvergne	(***Co-encadrant***)

Dédicaces

Je dédie ce travail,

A
L'âme de mes très chers grands parents, mon oncle

A
Mes chers parents

A
Mes chers frères et sœurs

A
Mes chers neveux et nièces

A
Toute ma grande famille et tous(tes) mes amis(es)

Dhouha GRISSA

Remerciements

Je tiens à remercier tout le monde, même si ces quelques mots ne puissent suffire pour exprimer tout ma gratitude envers toutes les personnes que j'ai eu l'honneur de rencontrer durant ces années de thèse.

Je remercie en premier lieu Monsieur Engelbert MEPHU NGUIFO, Professeur à l'Université Blaise-Pascal (France), Madame Sylvie GUILLAUME, Maître de Conférences à l'Université d'Auvergne (France) et Monsieur Sadok BEN YAHIA, Professeur à l'Université Tunis-El Manar, pour m'avoir proposé ce sujet, et m'avoir supporté dans toutes les étapes de cette thèse. Leurs avis, leurs nombreux conseils et leur soutien constant m'ont permis de m'améliorer sur de nombreux points. Qu'ils trouvent ici l'expression de mon plus grand respect et qu'ils reçoivent mes plus vifs remerciements.

Ce travail de thèse a été en partie réalisé dans la cadre d'un projet collaboratif, qui m'a permis d'explorer des domaines de recherche connexes à ma thèse et d'établir des collaborations avec d'autres chercheurs. J'adresse ainsi tout ma reconnaissance à Messieurs Radim BELOHLAVEK et Jan OUTRATA pour les échanges scientifiques et techniques que nous avons tenus et pour tout le temps qu'ils m'ont consacré.

J'adresse particulièrement mes remerciements aux membres du Jury. Je remercie Madame Amel BORGI et Monsieur Jean DIATTA, d'avoir accepté d'être rapporteurs de mes travaux. Je profite de l'occasion pour leur adresser mes sincères respects et leur exprimer ma profonde gratitude pour leurs commentaires et jugements très pertinents sur mon mémoire, tant sur le fond que sur la forme. La qualité de leurs travaux fait que leur point de vue m'est précieuse et je suis fière que mon travail soit soumis à leur jugement.

Je remercie également Messieurs Richard EMILION et Mohamed GAMMOUDI de m'avoir fait l'honneur de faire partie de mon jury de thèse.

Par la même occasion, je ne peux pas me retenir à remercier l'ensemble des thésards de m'avoir supporté tout au long de cette thèse. J'ai une pensée particulière pour ceux avec qui j'ai partagé de bons moments, notamment pour Ahlem Baccouche, Amani Kahloul, Audrey Lelong, Amina Aissani, Benoît Bernay, Fadi Zoubian, Haythem Touhami, Hayfa Naghmouchi, Hisham Requieg, Lakdhar Akroun, Nardjes Menadjelia, Pierre-Antoine Papon, Rim Douss, Sabeur Aridhi, Slim Bouker.

Je n'oublie jamais les enseignants, les chercheurs, le personnel du laboratoire LIMOS (UMR CNRS 6158, Clermont-Ferrand) et le laboratoire LIPAH (Tunis) je cite: Béatrice Bourdieu, Farouk Toumani, Faouzya, Martine Caccioppoli, Nicolas Champeil, Yannick Loiseau, Olivier Reynaud à qui j'adresse ma sincère gratitude et remerciement.

Cette thèse a été en partie financée par une bourse en alternance du Ministère de l'Enseignement Supérieur en Tunisie (M.E.S.) et également par le projet collaboratif Franco-Tunisien PHC Utique Exqui. Je tiens à remercier l'M.E.S. pour ses soutiens financiers, cette thèse n'aurait pu voir le jour sans cette aide.

La remise en question inévitable liée au travail de thèse m'a permis de me rendre compte de la chance que j'ai d'être si bien entourée. Je remercie mille fois mes parents et toute ma famille qui m'ont toujours soutenu dans tout ce que j'ai entrepris. Il m'est impossible d'imaginer en être arrivée là sans eux.

Mes remerciements vont aussi à tous mes amis, aussi étudiants et étudiantes de l'Université de Sousse, pour leur amitié et l'ambiance véritablement chaleureuse.

Les mots manquent aux émotions.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

La croissance des systèmes informatiques, notamment des capacités de stockage et de calcul, entraîne une explosion des volumes des données dans plusieurs secteurs d'activité tels que les télécommunications, la banque, la médecine, l'assurance, etc.. Le début des années 90 est alors marqué par l'émergence d'un nouveau domaine de recherche : l'extraction de connaissances à partir de données (*ECD ou plus couramment dénommée fouille de données*) consistant à exploiter et transformer des informations généralisables en connaissances nouvelles, valides, et potentiellement utiles à partir de grandes bases de données [FPSM91]. Les travaux de recherche effectués dans cette thèse s'inscrivent dans ce cadre.

Les connaissances dérivées à partir des données, représentent en fait de véritables mines d'or pour les experts. Pour les découvrir, de nombreuses techniques ont été utilisées parmi lesquelles la recherche de règles d'associations qui a connu un fabuleux essor depuis le début des années 90. Introduits par Agrawal et al. [AIS93], les algorithmes de découverte de règles d'association (*voir [HGN00] pour une synthèse*) reposent classiquement sur ce couple de mesures : le support et la confiance. Étant donnés un seuil minimal de support et un seuil minimal de confiance, il s'agit de produire l'ensemble de toutes les règles possibles ayant des valeurs du support et de la confiance supérieures aux seuils fixés au préalable par l'utilisateur. Toutefois, la quantité de règles générées par ces algorithmes est prohibitive, essentiellement lorsque les valeurs des seuils est faible. Elle croît exponentiellement avec le nombre d'attributs qui décrit les données.

Les deux mesures *support* et *confiance* sur lesquelles sont fondés les algorithmes d'extraction de règles d'association, sont insuffisantes pour garantir la qualité des règles révélées. Elles ont été remises en cause dans de nombreux travaux comme par exemple [SM02]. En pratique, une étape de post-traitement s'avère nécessaire pour découvrir des connaissances intéressantes à partir de l'ensemble de règles obtenues à la sortie des algorithmes. Afin de franchir ces problèmes, le premier de nature quantitatif (*beaucoup de règles extraites*) et le second de nature qualitatif (*beaucoup de règles redondantes et non pertinentes*), différentes solutions ont été proposées. L'une des solutions consiste à restituer facilement et de manière synthétique l'information extraite grâce à des techniques de représentation visuelle [HW01].

Une autre consiste à réduire le nombre de règles extraites. Certains auteurs [Zak00], [ZTS04] éliminent les règles redondantes, d'autres évaluent et ordonnent les règles extraites grâce à d'autres mesures d'intérêt [LMPV03], [TGCTB13], autres que les mesures support

et confiance. Néanmoins, nombreuses sont les mesures d'intérêt [PS91b], [ST96], [BMS97], [Fre99], [HH01], [TKS02], [LMPV03], [LMP+03], [BKGB04], [TKS04], [McG05], qui ont été proposées dans la littérature. Comme l'intérêt dépend à la fois des préférences de l'utilisateur et des données, les mesures ont été répertoriées en deux catégories [ST96], [Fre99], [GH07] : les mesures subjectives (*orientées utilisateur*) et les mesures objectives (*orientées données*) [SG91].

Dans le présent travail, nous nous intéressons aux mesures objectives, qui permettent de filtrer les règles et de les ordonner selon leur intérêt, ignorant par conséquent les règles de faible qualité. Néanmoins, un nombre important de mesures d'intérêt existe dans la littérature induisant à son tour de nouveaux problèmes, tel que le problème de sélection de mesure d'intérêt qui soit la mieux appropriée aux besoins de l'utilisateur expert. Généralement, il est difficile de choisir les mesures à appliquer. D'où, face à cette difficulté de sélection de mesures, plusieurs travaux ont été réalisés selon des points de vue différents [PS91a], [BMS97], [Fre99], [LMP+03], [LT04], [HGB+07], [JBC13], [TGCTB13] afin de trouver une solution adéquate à ce problème. Dans cette thèse, nous proposons une étude approfondie du comportement d'une soixantaine de mesures d'intérêt existantes dans la littérature, par une approche par catégorisation et selon deux points de vue : théorique et empirique.

Contributions de la thèse

Avec le nombre important de mesures d'intérêt existant dans la littérature pour évaluer les règles d'association, des études comparatives ont été réalisées pour aider l'utilisateur dans le choix de la mesure la mieux appropriée à ses attentes. C'est dans le cadre de cette problématique que se situe ces travaux de thèse, où nous proposons une étude approfondie du comportement d'une soixantaine de mesures d'intérêt.

Les contributions de cette thèse sont les suivantes :

1. **Étude formelle des mesures d'intérêt**
   Tout d'abord, nous étudions le comportement des mesures d'intérêt selon des propriétés formelles. Pour ce faire, nous recensons dans la littérature une vingtaine de propriétés des mesures, que nous formalisons pour éviter toute ambiguité sur celles-ci. Nous évaluons par la suite l'ensemble des mesures selon ces propriétés formelles. Sur ce point de la thèse, il en résulte :
   – une synthèse des différentes propriétés et mesures existants dans la littérature ;
   – une formalisation des propriétés de mesures ;
   – une matrice d'évaluation décrivant le comportement de 61 mesures selon 19 propriétés.

2. **Classification des mesures d'intérêt**

   Afin d'aider l'utilisateur dans le choix de mesures d'intérêt qui soient les mieux appropriées à ses besoins, nous effectuons une classification d'une soixantaine de mesures d'intérêt existants dans la littérature. D'où, deux types de méthodes de regroupement sont appliquées : le premier constitue la méthode de classification hiérarchique et celle de partitionnement des *k-moyennes* et produit des groupes ou classes disjointes de mesures ; le deuxième, complémentaire du premier, produit quant à lui des classes recouvrantes, et la méthode appliquée est l'analyse factorielle booléenne. Au niveau de ce deuxième point de la thèse, il en résulte :

   – 8 classes de mesures selon la classification hiérarchique et aussi selon la méthode des *k-moyennes* ;
   – 8 classes fortes de mesures suite à une dizaine de répétitions de l'algorithme des *k-moyennes* ;
   – un consensus sur la classification, où 7 classes de mesure en résulte ;
   – une interprétation des 7 classes de mesures ;
   – la proposition de mesures référentes pour chaque classe ;
   – l'identification de classes recouvrantes et leur confrontation aux résultats de la classification disjointe ;
   – une étude comparative avec des travaux existants dans la littérature sur la classification des mesures.

3. **Étude expérimentale des mesures d'intérêt**

   Nous effectuons une étude empirique du comportement des mesures d'intérêt dans le but de valider les résultats de la classification formelle. Cette étude est réalisée sur des jeux de données avec des caractéristiques différentes, elle nous permet d'identifier différentes catégories de mesures d'intérêt. À l'issue de ce dernier point de la thèse, il en résulte :

   – 3 catégories de mesures :
     - des mesures au comportement semblable, i.e., les mesures proposent pratiquement les mêmes *N* meilleures règles ;
     - des mesures au comportement différent, i.e., les ensembles des *N* meilleures règles proposés par les mesures sont différents ;
     - des mesures dont le comportement est en fonction des données, i.e., dont le comportement varie selon le jeu de données.
   – une interprétation des groupes de mesures stables, qui appartiennent à la première catégorie ;
   – une confrontation de ces groupes stables aux 7 classes de mesures formelles ;

– une étude comparative avec des travaux de la littérature.

Structure de la thèse

Cette thèse est constituée de cinq chapitres.

Le premier chapitre concerne les notions de base, considérées utiles dans le domaine d'extraction de connaissances. Il comprend un bref résumé des différentes étapes du processus d'ECD, les états de l'art sur les notions des règles d'association (*type de connaissances sur quoi porte notre étude*), les algorithmes d'extraction de règles type Apriori permettant l'identification de telles règles, et les mesures d'intérêt capables d'évaluer la qualité de ces règles ainsi que de les ordonner. Il passe également en revue les différents travaux réalisés sur les mesures d'intérêt de règles d'association de point de vue formel et empirique.

Le deuxième chapitre intitulé "*Étude formelle des mesures*", présente tout d'abord quelques exemples de mesures d'intérêt objectives définies dans la littérature. Il décrit ensuite une vingtaine de propriétés formelles proposées dans la littérature afin de dévoiler les qualités d'une *bonne* mesure d'intérêt. Une formalisation de ces propriétés est proposée pour éviter toute ambiguité sur celles-ci. Ce chapitre se termine par une étude formelle des mesures, qui consiste en la construction d'une matrice évaluant une soixantaine de mesures d'intérêt sur les propriétés formelles retenues.

Le troisième chapitre intitulé "*Classification des mesures : méthodes sans recouvrement*", concerne la catégorisation des mesures. Il dresse dans un premier temps un panorama des différentes approches proposées pour regrouper des données. Il présente essentiellement les méthodes hiérarchiques et de partitionnement. Ces dernières seront appliquées par la suite sur la matrice d'évaluation des mesures selon les propriétés, illustrée dans le deuxième chapitre, pour construire un ensemble de groupes disjoints de mesures. Une étude comparative avec les travaux existants dans la littérature sur la classification des mesures d'un point de vue formel est réalisée à la fin de ce chapitre.

Cette dernière classification est complétée dans le *chapitre 4* par une deuxième catégorisation des mesures, qui permet de construire des groupes recouvrants. Le début de ce chapitre dresse en bref les méthodes de classification avec recouvrement et s'intéresse plus particulièrement à la méthode d'analyse factorielle booléenne. Ayant les mêmes données en entrée que celles utilisées lors du *chapitre 3*, il s'agit d'appliquer l'analyse factorielle booléenne pour obtenir des groupes recouvrants de mesures, qui vont être interprétés et comparés par la suite. Ce chapitre est clos par une discussion des résultats de la classification obtenus par les deux méthodes sans et avec recouvrement. Cette discussion entraîne l'obtention de groupes stables de mesures d'intérêt, i.e., de mesures qui en confontant les résultats de la classification, sont

toujours regroupées ensemble indépendemment de la méthode appliquée.

Si des études approfondies ont été réalisées sur les mesures d'intérêt d'un point de vue formel, nous jugeons essentiel l'évaluation des mesures selon les données afin de voir si les mesures d'un même groupe proposent les mêmes règles d'association, c'est l'objet du dernier chapitre de cette thèse, intitulé "*Étude expérimentale des mesures d'intérêt*". Cette étude complémentaire effectuée sur les mesures d'un point de vue analyse de données tient compte de la nature et du type des données analysées. Elle propose alors un cadre qui consiste à comparer les ensembles de *N* meilleures règles fournis par chaque mesure. Cette comparaison va aboutir à l'obtention de 3 catégories de mesures suivantes :

- des mesures au comportement semblable ;
- des mesures au comportement différent ;
- des mesures dont les comportement est en fonction des données.

Nous nous intéressons essentiellement aux mesures de la première catégorie. Ces groupes de mesures au comportement similaire vont être interprétés et confrontés aux groupes de mesures obtenus par l'étude formelle, et seront comparés par la suite avec des travaux de la littérature.

Finalement, nous concluons notre travail et dressons quelques perspectives.

CHAPITRE 1

État de l'art

Sommaire

1.1 Introduction

Ce travail s'inscrit dans le domaine de l'*Extraction de Connaissances à partir des Données* (*ECD*), appelé *Knowledge Discovery in Databases (KDD)* en anglais. Dans le contexte de ce

domaine, nous explorons l'un des principaux problèmes, à savoir la recherche de règles d'association. Ce problème majeur en ECD permet, à partir de bases de données volumineuses, la génération d'un très grand nombre de règles difficilement exploitables par l'utilisateur. Il s'avère donc nécessaire de mettre en oeuvre une étape supplémentaire de post-traitement pour aider l'utilisateur à réduire le nombre de règles découvertes. Ainsi, l'une des solutions envisagées est la proposition de mesures d'intérêt d'extraction de règles d'association [ST96], [McG05]. Notre objectif est alors de mener une étude approfondie du comportement de ces mesures selon deux points de vue : théorique et expérimental dans le but de répondre aux besoins de l'utilisateur final, à savoir la validation et la sélection des règles pertinentes.

Ce chapitre est organisé de la manière suivante : nous commençons par définir brièvement le processus d'extraction de connaissances à partir des données. Nous présentons par la suite le domaine d'étude, les notations utilisées et les règles d'association. Puis, nous abordons l'étape d'extraction de règles d'association dans laquelle nous introduisons les algorithmes d'extraction des règles. Dans la *section 1.4*, nous abordons les méthodes d'évaluation des règles d'association et plus précisément les mesures d'intérêt. Enfin, nous proposons une revue critique de la littérature en présentant les différents travaux réalisés sur les mesures d'intérêt selon un cadre formel et empirique avant de terminer par les limites de l'existant et la description de notre motivation.

1.2 Extraction de connaissances à partir des données

Le domaine d'*Extraction de Connaissances à partir des Données* (ECD), est un domaine de recherche complexe et interactif, introduit en *1991* lors du premier workshop KDD animé par Piatetsky-Shapiro [PS91b]. Poussé par le secteur industriel et fondé sur l'analyse des données, les statistiques et l'apprentissage automatique, l'ECD a suscité l'intérêt de nombreux chercheurs [FPSM91], [BA96], [FPSSU96], [KS96], [HMS01]. Plusieurs définitions de l'ECD ont été proposées dans la littérature. Frawley et al. [FPSM91] le définissent comme "*un processus non trivial consistant à extraire, à partir de l'ensemble des données, une information implicite, inconnue auparavant et potentiellement utile*", *"Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.*". Fayyad et al. [FPSSU96] le spécifient comme étant "*un processus non trivial pour l'identification d'une connaissance valide, nouvelle, potentiellement utile et éventuellement compréhensible à partir des bases de données*", "*the nontrivial process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases*".

Ainsi, l'objectif de l'ECD est de faire face à l'émergence de données volumineuses par la découverte de ces connaissances qui doivent être à la fois "pertinentes", "intéressantes" et "utiles"

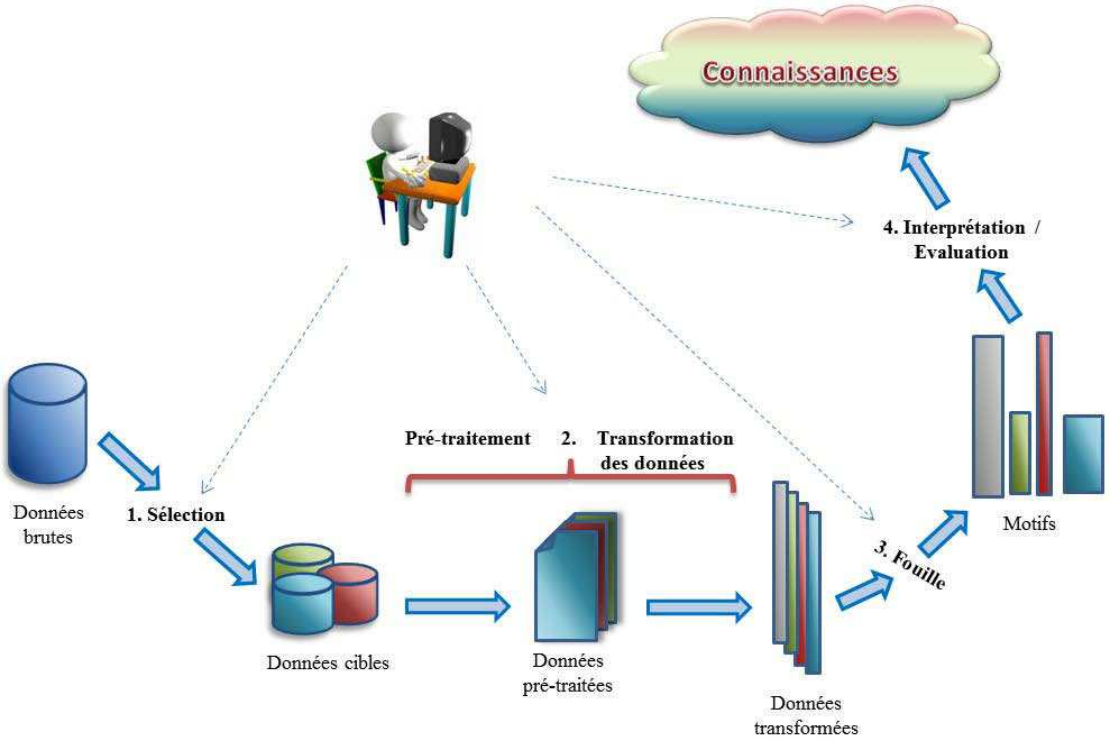

FIGURE *1.1: Les étapes du processus ECD.*

pour l'utilisateur. C'est en effet un processus interactif et itératif avec lequel l'utilisateur interagit pour prendre les décisions adaptées à ses besoins. Le processus de l'ECD est constitué de quatre phases principales allant de la sélection et la préparation des données (*pré-traitement*) jusqu'à l'interprétation et l'évaluation des données (*post-traitement*), en passant par la phase de recherche d'informations (*la fouille des données*). Ces quatre phases sont illustrées dans la *figure 1.1* [FPSSU96] et développées dans ce qui suit.

1. **La sélection** : cette première étape du processus consiste à garder parmi l'ensemble des données, qu'une sous-partie intéressante par rapport au problème. Ainsi, il s'agirait en effet d'identifier les informations adaptées aux besoins de l'utilisateur pour une application bien déterminée.

2. **Le pré-traitement et transformation des variables** : le rôle de cette deuxième étape du processus est de préparer les données précédemment sélectionnées, pour les adapter à l'étape de fouille. Un traitement informatique sera ainsi appliqué aux informations conservées. Il s'agit par la suite de présenter les données traitées sous la forme imposée par la méthode de fouille retenue (*algorithme d'extraction de connaissances*). Ces données peuvent être présentées sous forme de variables binaires, discrètes ou continues selon la méthode retenue.

3. **La fouille** : cette troisième étape est au coeur du processus de l'ECD puisqu'elle permet d'identifier et de mettre en évidence des informations ou des connaissances à partir de données transformées. Les informations ainsi générées peuvent prendre différentes formes selon la méthode utilisée et le problème à résoudre : tels que les arbres de décisions (*l'analyse prédictive*) ou les règles d'association (*l'analyse exploratoire*).
4. **L'interprétation et l'évaluation des informations** : une quantité importante d'informations peut être générée à partir des algorithmes d'extraction des données, dont la plupart sont inutiles ou redondantes. Pour pallier ce défaut, une dernière étape de post-traitement des connaissances découvertes s'avère indispensable pour transformer ces connaissances extraites en connaissances intéressantes et facilement exploitables par l'utilisateur. Il serait donc intéressant d'appliquer un filtrage automatique sur l'ensemble des informations extraites et de les ordonner afin de faciliter la prise de décision et l'interprétation des résultats révélés à l'utilisateur. Dans la littérature, il existe plusieurs moyens pour assister l'utilisateur dans son travail, comme l'utilisation de mesures d'intérêt (*définies dans la section 1.5*) qui permettraient d'évaluer la qualité des connaissances extraites.

Dans ce qui suit, nous allons nous intéresser aux règles d'association qui visent à découvrir des tendances, a priori inconnues, au sein des données.

1.3 Règles d'association

Dans cette section, nous commençons par présenter le domaine de recherche. Par la suite, nous définissons les notations utilisées et les règles d'association.

1.3.1 Présentation

Nous présentons un domaine de recherche assez mature en fouille de données [CR06], celui de l'extraction de règles d'association. Ce thème fut introduit par "Agrawal et al.," dans [AIS93] dans le but d'analyser les bases de données transactionnelles pour découvrir des habitudes d'achat des clients dans un supermarché, comme les articles qui sont souvent achetés ensemble. Connu sous le nom de *panier du consommateur*, ce problème est très courant en ECD et il a reçu beaucoup d'attention de la part des chercheurs [AIS93], [AS94], [KMR+94]. Il s'agit de découvrir parmi les ensembles des transactions [1], un ensemble de règles qui exprime une possibilité d'association entre différents articles. L'extraction des règles d'association est

1. Une transaction est constituée de l'ensemble des articles ou éléments achetés par un client.

TID	Lait	Banane	Café	Pizza	Sucre
1	1	0	0	1	0
2	1	1	1	0	1
3	1	0	1	0	0
4	0	1	0	1	0
5	1	1	1	1	0

TABLE *1.1: Représentation binaire des données de "paniers" de clients.*

ainsi une technique qui "*vise à extraire des corrélations intéressantes, des modèles fréquents ou des associations entre les ensembles d'éléments dans les bases de données de transaction ou autres dépôts de données* " [KK06] (p.71).

On se situe dans le cadre d'une base de données D binaire, définie par un ensemble $I = \{i_1, i_2, \ldots, i_p\}$ de p attributs également appelés *items* et un ensemble $T = \{t_1, t_2, \ldots, t_n\}$ de n éléments, reliés par une relation binaire R. Chaque transaction t_i désigne donc un sous-ensemble de I, ayant un identifiant *TID* (*Transaction IDentifier*). Un sous-ensemble de I est appelé *itemset* ou *motif*.

La *table 1.1* illustre ce type de données en restituant un exemple de 5 paniers d'achats. Nous sommes donc en présence d'un ensemble T de 5 transactions $\{t_1, \ldots, t_5\}$, décrites par 5 items $\{i_1 =$ *Lait,* $i_2 =$ *Banane,* $i_3 =$ *Café,* $i_4 =$ *Pizza,* $i_5 =$ *Sucre*$\}$ ou articles achetés par les clients.

Le jeu de données de la *table 1.1* sera utilisé par la suite pour décrire, en cas de besoin, les notions introduites tout au long de ce manuscrit.

Dans ce qui suit, nous présentons les notations usuelles des règles d'association avant de les définir.

1.3.2 Notations usuelles

Dans la suite, nous utilisons les notations suivantes :

- T représente la base de données ;
- I représente l'ensemble des attributs de T ;
- T_X représente l'ensemble des transactions qui contiennent le motif X ou encore tous les items i_j ($j \in \{1, \ldots, p\}$), contenus dans le motif X ;
- $n = |T|$ représente le nombre d'enregistrements ou de transactions de la base ;
- $n_X = |(X)_{t_i \in T}| = |T_X|$ représente le nombre de transactions satisfaisant le motif X ;
- $n_{XY} = |T_{X \cup Y}| = |T_{XY}|$ représente le nombre de transactions satisfaisant à la fois X et Y ;
- $n_{\overline{X}} = n - n_X$, où $\overline{X}$ représente la négation de X ;

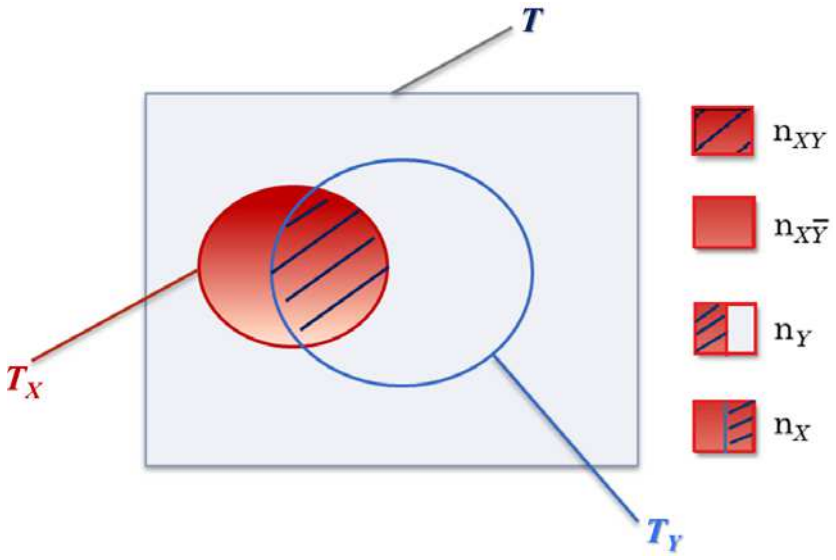

FIGURE *1.2: Notations utilisées pour une règle* $X \rightarrow Y$.

– $n_{X\overline{Y}} = |T_{X \cup \overline{Y}}| = |T_{X\overline{Y}}|$ représente le nombre de transactions satisfaisant X mais pas Y.

Le diagramme de Venn de la *figure 1.2* illustre ces différentes notations.

Pour plus de clarté, nous gardons aussi les notations probabilistes : $P(X)$ (*resp.* $P(XY)$, $P(X\overline{Y})$) comme étant la probabilité de X (*resp.* Y, XY, $X\overline{Y}$). Cette notation est équivalente à celle décrite ci-dessus puisqu'il s'agit de passer de l'une à l'autre simplement via la relation $P(X) = \frac{n_X}{n}$. Selon nos besoins, nous gardons l'une de ces notations, celle qui sera la plus appropriée.

Ayant introduit ces différentes notations, nous pouvons maintenant définir une règle d'association.

1.3.3 Définition

Définition 1 ***(règle d'association)*** *Une règle d'association [AIS93] est un couple* (X, Y)*, noté* $X \rightarrow Y$*, où* X *et* Y *sont des motifs (ou conjonctions de variables binaires) disjoints. Nous avons donc les relations suivantes :* $X \subseteq I$, $Y \subseteq I$ *et* $X \cap Y = \varnothing$.

Une règle d'association de type $X \rightarrow Y$ prend la forme "*Si condition alors résultat*". Elle comporte une partie prémisse (*ou antécédent*) composée d'un ensemble d'items X et une partie conclusion (*ou conséquent*) composée d'un ensemble d'items Y disjoint de X. Une telle règle permet de découvrir si les transactions qui vérifient le motif X ont tendance à vérifier également le motif Y. Un exemple de règle qui pourrait être extraite des données de la *table 1.1*, est "$Café \rightarrow Lait$".

En outre, une règle d'association est entièrement caractérisée par son tableau de contingence (*table 1.2*), qui est la base pour le calcul des mesures d'évaluation des règles d'association. Les cellules du tableau de contingence sont liées par les relations suivantes, où $P(XY)$ est exprimée comme le rapport $\frac{n_{XY}}{n}$:

- $P(X) + P(\overline{X}) = 1$
- $P(XY) = P(Y) - P(\overline{X}Y) = P(X) - P(X\overline{Y})$

	Y	$\overline{Y}$	$Profils\ ligne$
X	$P(XY)$	$P(X\overline{Y})$	$P(X)$
$\overline{X}$	$P(\overline{X}Y)$	$P(\overline{X}\overline{Y})$	$P(\overline{X})$
$Profils\ colonne$	$P(Y)$	$P(\overline{Y})$	1

TABLE *1.2: Tableau de contingence.*

Les règles d'association ont été utilisées avec succès dans une large variété de domaines d'application [Hue09], parmi lesquels le secteur médical pour la recherche par exemple de complications dues à des associations de médicaments [MYGS91], [OO98], [PMS97], [GSM94], l'analyse d'images [Czy96], [OO98], [ZHL+98], de données génomiques [XHD+05], de données graphiques [KK04] et statistiques [SW85], l'amélioration des services de télécommunications [HKM+96], [KMT97], [AMS97], la fouille de textes [Kod99], [HY02]. Cette technique peut aussi être appliquée à tout autre secteur d'activité où il serait intéressant de découvrir des conjonctions d'articles ou services qui apparaissent fréquemment ensemble tels que les services bancaires.

Pour évaluer la qualité des règles d'association extraites, deux mesures sont classiquement utilisées : le support et la confiance qui sont l'objet de notre prochaine section.

1.3.4 L'approche support-confiance

Dans ce qui suit, nous définissons les deux mesures les plus utilisées, *support* et *confiance*, proposées par [AIS93], [AS94].

Définition 2 ***(support)*** *Le support d'une règle $X \rightarrow Y$ désigne la proportion de transactions qui vérifient à la fois X et Y, c'est-à-dire la fréquence d'apparition de X et Y.*

$$support(X \rightarrow Y) = P(XY) = \frac{|T_{XY}|}{|T|} = \frac{n_{XY}}{n} \tag{1.1}$$

Le support est ainsi le rapport du nombre d'enregistrements où la prémisse et la conclusion sont vérifiées, sur le nombre total d'enregistrements.

Par exemple, dans les données du panier de la *table 1.1*, nous avons :

$$support(Lait) = P(Lait) = \frac{4}{5} = 80\%$$
$$support(Lait \rightarrow Café) = P(Lait, Café) = \frac{3}{5} = 60\%$$

Définition 3 ***(confiance)*** *La confiance d'une règle $X \rightarrow Y$ est la proportion de transactions qui vérifient Y parmi celles qui réalisent X, c'est-à-dire la fréquence conditionnelle de Y sachant X.*

$$confiance(X \rightarrow Y) = P(Y|X) = \frac{P(XY)}{P(X)} = \frac{|T_{XY}|}{|T_X|} \tag{1.2}$$

La confiance est ainsi le rapport du nombre d'enregistrements où la prémisse et la conclusion sont vérifiées, sur le nombre d'enregistrements où seule la prémisse est vérifiée. Elle ne tient pas compte du nombre total d'enregistrements n, elle est uniquement fonction de $\mid T_{XY} \mid$ et $\mid T_X \mid$ et pas de la taille de la base $\mid T \mid$.

Afin de retenir les règles les plus intéressantes, que l'on appelle *règles valides*, des seuils d'élagage doivent être fixés. Ces deux seuils, notés respectivement "min_{sup}" pour le support minimum et "min_{conf}" pour la confiance minimum, ont pour objectif d'éliminer les règles les moins intéressantes. Le choix des seuils s'avère essentiel pour que le nombre de règles à proposer à l'utilisateur soit gérable. Reprenons l'exemple du panier de la *table 1.1, page 11*, issu d'une base de données transactionnelle d'un supermarché, où la règle suivante $Pizza \rightarrow Lait$ a été déduite. Cette règle est générée avec un support égal à 40% (*i.e., dans 40% des transactions de la base de données, contiennent les items $Pizza$ et $Lait$*) et une confiance de 67% (*i.e., 67% des fois lorsque le consommateur achète la pizza alors il achète également du lait*). Une telle règle est jugée *valide* si et seulement si son support et sa confiance sont supérieurs ou égaux aux seuils respectifs min_{sup} et min_{conf} préalablement fixés par l'utilisateur. Pour le cas de la règle $Pizza \rightarrow Lait$, si nous retenons des valeurs supérieures à 40% pour min_{sup} et à 67% pour min_{conf}, cette règle sera alors ignorée. Afin de choisir les bons seuils et de garder les règles intéressantes, il est préférable de sélectionner de petites valeurs pour le support et de fortes pour la confiance.

Ayant présenté l'approche support-confiance, nous abordons dans ce qui suit le problème d'extraction de règles d'association.

1.4 Extraction de règles d'association

Le problème d'extraction de règles d'association est l'une des 4 principales étapes de l'ECD (*décrits brièvement dans la section 1.2*), qui vise à découvrir des liaisons significatives entre les items dans une base de données. Il peut être formulé selon les contraintes suivantes :

Étant donné un ensemble de transactions T, trouver toutes les règles d'association $X \rightarrow Y$ ayant un support supérieur ou égal à min_{sup} et une confiance supérieure ou égale à min_{conf}, où min_{sup} et min_{conf} sont deux seuils minimums pour le support et la confiance fixés par l'utilisateur.

L'extraction de règles d'association consiste ainsi à déterminer l'ensemble des règles, dont le support et la confiance sont au moins égaux à des seuils minimaux de support et de confiance fixés par l'utilisateur. En se basant sur la contrainte du support, il est possible de réduire le nombre des motifs proposés [AIS93], et ne garder que les motifs fréquents que nous définissons dans ce qui suit.

1.4.1 Recherche des motifs fréquents

La recherche de motifs fréquents dans les bases de données fait l'objet, depuis quelques années, de recherches intensives dans le domaine de la fouille de données. Cette phase est primordiale dans l'extraction des règles d'association, et elle consiste à rechercher des motifs ou des associations de variables que l'on rencontre fréquemment ensemble.

Définition 4 ***(motif fréquent)*** *Un motif X est fréquent lorsque la valeur de son support $support(X \rightarrow Y)$ est supérieure au seuil minimum min_{sup} fixé par l'utilisateur : $support(X \rightarrow Y) \geq min_{sup}$ [AS94].*

Si nous choisissons un seuil minimum égal à 30% ($min_{sup} = 30\%$) et que nous l'appliquons sur les données du panier (*table 1.1*), nous obtenons la liste des motifs fréquents suivants : {*Lait*}, {*Café*}, {*Banane*}, {*Pizza*}, {*Lait, Café*}, {*Lait, Banane*}, {*Café, Banane*}, {*Pizza, Banane*} et {*Lait, Banane, Café*} sont tous des motifs fréquents. Les motifs {*Sucre*}, {*Sucre, Banane*} ou encore {*Sucre, Café*}, dont le support est inférieur à 30%, ne sont pas fréquents.

Le problème de l'extraction des motifs fréquents est de complexité exponentielle dans la taille n de l'ensemble d'items puisque le nombre de motifs fréquents potentiels est 2^n.

Afin de réduire l'espace de recherche des motifs fréquents, les algorithmes d'extraction de règles d'association reposent sur la propriété d'anti-monotonie.

Propriété 1 ***(anti-monotonie)*** *Soient X et Y deux motifs disjoints. Nous avons (dans le cas de support) :*

$$\forall X, Y \subseteq I \ : X \subseteq Y \Rightarrow support(X) \geq support(Y).$$

La *propriété 1.4.1* [AS94], [MTV94] est particulièrement importante dans les algorithmes d'extraction de connaissances, puisqu'elle permet d'affirmer que pour un motif $X \subseteq I$:

(i). Si X est fréquent, alors pour tout motif X_1, tel que $X_1 \subseteq X$, X_1 est aussi fréquent, *i.e., Tout sous-ensemble d'un motif fréquent est fréquent.*

(ii). Si X est non fréquent, alors pour tout motif X_2, tel que $X \subseteq X_2$, X_2 est aussi non fréquent, *i.e., Tout sur-ensemble d'un motif non fréquent est non fréquent.* (*anti-monotonie*)

Plusieurs travaux se sont intéressés à la génération des motifs fréquents, et à la découverte de toutes les règles d'association valides liant ces motifs entre eux.

Ce problème de découverte des règles associatives [AIS93] peut ainsi être décomposé en deux sous-problèmes :

1. **Trouver tous les motifs fréquents** ;
2. **Générer l'ensemble des règles associatives**, ayant une grande confiance supérieure à min_{conf} à partir des motifs fréquents. Ces règles sont appelées règles valides.

Parmi les algorithmes fondateurs pour la recherche de règles associatives, nous retrouvons Apriori [AMS+96]. Cet algorithme se base essentiellement sur la propriété d'anti-monotonicité de Support existant entre les motifs. Il permet d'évaluer les règles potentiellement valides, et de ne garder que celles qui satisfont les mesures d'intérêt *support* et *confiance*.

1.4.2 Principe de l'algorithme Apriori

L'algorithme *Apriori* est l'un des plus importants algorithmes d'extraction de règles d'association [WKQ+07], basé sur l'approche support-confiance. Comme il est fondé sur la propriété d'anti-monotonie, *Apriori* est alors capable d'élaguer les motifs non fréquents d'une base de données volumineuse. Pour ce faire, il s'appuie sur le treillis des motifs.

Définition 5 ***(Treillis)*** *Un ensemble ordonné $(Tr, \preceq)$ est un treillis si toute paire d'éléments de Tr possède une borne inférieure et une borne supérieure.*

Un exemple de treillis est illustré dans la *figure 1.3*. Ce treillis des motifs représente les données du "panier" (*décrites dans la table 1.1*) sous forme de diagramme de Hasse. C'est

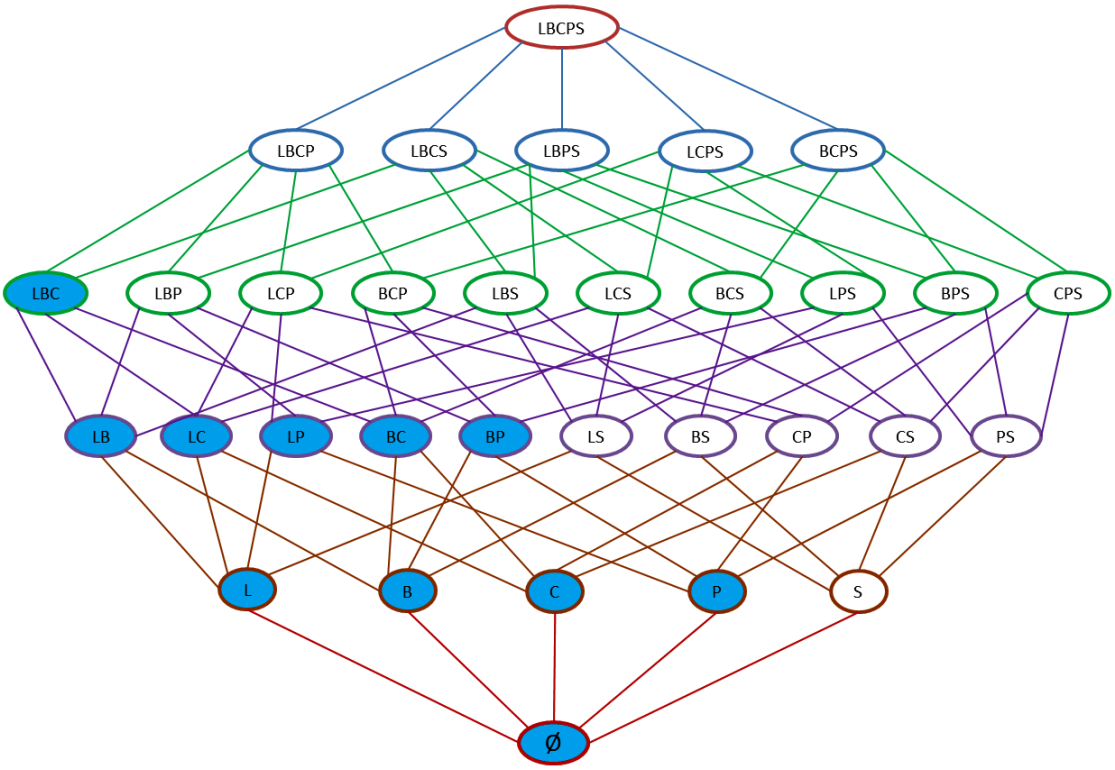

FIGURE *1.3: Exemple de treillis associé à un ensemble de 5 items.*

en effet une représentation visuelle d'un ordre fini de l'ensemble des motifs selon la relation d'inclusion ensembliste. Nous désignons par les lettres *L, B, C, P* et *S*, respectivement les articles *Lait, Banane, Café, Pizza* et *Sucre*.

Pour un ensemble I composé de $p = 5$ items, le treillis contient 32 motifs (2^p) et sa hauteur est égale à 6 (*p+1*). Le parcours de ce treillis se fait par niveaux du bas (*où le motif est vide*) vers le haut (*où on a tous les motifs*) en déterminant à chaque itération k tous les motifs fréquents de taille k. La *figure 1.3* montre des couleurs différentes pour chaque niveau et représente les motifs fréquents par les noeuds ayant un fond bleu. Néanmoins, afin d'éviter de parcourir un même noeud ou motif du treillis plusieurs fois, les auteurs dans [AIS93] définissent un ordre total $\prec$ sur les motifs, qui est l'*ordre lexicographique*[2]. Cet ordre semble nécessaire afin de limiter l'espace de recherche pour trouver les motifs fréquents, en garantissant que chaque noeud du treillis soit parcouru une seule fois par l'algorithme.

Dans le cas où un motif non fréquent se présente, alors tous les motifs situés au dessus, i.e., tous les sur-ensembles de motifs sont aussi non fréquents et par conséquent il est inutile de calculer leurs supports. Cette propriété d'anti-monotonie va permettre ainsi, pour chaque

2. Un ordre lexicographique est une relation d'ordre sur E^k, où E est un ensemble totalement ordonné et k un entier. On la définit de la manière suivante : $(x_1, x_2, \cdot, x_k) \leq (y_1, y_2, \cdot, y_k)$, si et seulement si il existe i tel que pour tout $j < i$, $x_i \leq y_i$

Notations	
min_{sup}	Seuil minimum du support
IF	Ensemble des motifs fréquents
C_k	Ensemble des k-motifs candidats
F_k	Ensemble des k-motifs fréquents
k	Numéro de l'itération et taille du motif
C_{i_p}	Ensemble des candidats possédant le motif i_p

TABLE *1.3: Notations utilisées dans l'algorithme Apriori*

itération k, de générer l'ensemble des k-motifs candidats (*i.e., tous les motifs potentiellement fréquents de taille* k) et d'élaguer les candidats non fréquents. L'algorithme recherche donc itérativement les ensembles de $(k+1)$-motifs fréquents à partir des ensembles fréquents de cardinal k déterminés lors de l'itération précédente.

Nous expliquons dans ce qui suit le pseudo-code de l'algorithme Apriori d'Agrawal et Srikant [AS94].

1.4.3 Algorithme Apriori

Le pseudo-code de l'algorithme *Apriori* est présenté dans l'*algorithme 1* et les notations que nous utilisons sont présentées dans la *table 1.3*.

Algorithme 1: Algorithme de génération des motifs fréquents

Input : Base de données $\mathbb{T}$, seuil minimum du support min_{sup}
Output : Ensemble des motifs fréquents *IF*

```
1  Apriori-Gen
2  begin
3      Calculer F_1
4      k ← 2
5      for k ; F_{k-1} ≠ ∅ ; k++ do
6          C_k ← Apriori − Gen(F_{k-1})
7          for chaque item i_p de I do
8              C_{i_p} ← Subset(C_k, i_p)
9              for chaque candidat C ∈ C_{i_p} do
10                 support(C).count++
11         F_k ← {C ∈ C_k / support(C) ≥ min_sup}
12     Retourner IF = ∪_k F_k
```

L'*algorithme 1* ainsi introduit permet de découvrir les motifs fréquents en partant de ceux de taille 1 (*ligne 3*), on note cet ensemble F_1. Chaque itération k (*lignes 4 à 12*) se subdivise en

deux étapes :

1. La première étape fait appel à la procédure $Apriori - Gen$.

   Apriori-Gen Cette procédure, décrite dans l'*algorithme 2*, est aussi constituée de deux phases :

   (a) La première phase nommée *Joindre* (*lignes 2 à 5 de l'algorithme 2*) permet de déterminer l'ensemble C_k des k-motifs candidats, i.e., les k-motifs qui sont potentiellement fréquents à partir des $(k-1)$-motifs fréquents de F_{k-1}.

   (b) La deuxième phase nommée *Effacer* (*ligne 6 à 9 de l'algorithme 2*) consiste à supprimer de C_k les éléments qui ne vérifient pas la *propriété d'anti-monotonie (i)* (*cf. section 1.4.1*) des sous-ensembles fréquents. Deux motifs X et Y de F_{k-1} forment un motif C si, et seulement s'ils ont $(k-2)$ attributs (*dans le prefixe*) en commun, ce qui est exprimé en utilisant l'ordre lexicographique (*ligne 5*).

2. La deuxième étape (*lignes 7 à 11 de l'algorithme 1*) fait appel à la procédure $Subset(C_k, i_p)$.

   Subset Cette procédure détermine les k-motifs fréquents parmi les k-motifs candidats. Il s'agit de trouver pour chaque motif i_p de l'ensemble I (*ligne 7*), l'ensemble C_{i_p} des k-motifs candidats qu'il possède (*au moyen de la procédure* $Subset$*, ligne 8*). Ainsi, pour trouver le support de chaque candidat, il s'agit de parcourir la base de données. Une fois l'ensemble C_{i_p} déterminé, le support du candidat sera incrémenté (*ligne 10*). Parmi les candidats, seuls ceux qui ont le support supérieur à min_{sup} sont retenus (*ligne 11*). Le processus s'arrête lorsqu'aucun nouveau motif candidat ne peut être généré, i.e., lorsque $F_{k-1} = \varnothing$.

Disposant des motifs fréquents, il nous faut maintenant découvrir les règles d'association.

1.4.4 Génération des règles d'association

Durant cette étape, il s'agit de générer l'ensemble R des règles d'association à partir de l'ensemble IF de motifs fréquents trouvés durant la phase précédente (*trouver les motifs fréquents*). Pour chaque motif fréquent X, nous considérons tous ses sous-ensembles (*d'après la propriété d'antimonotonicité, sont tous fréquents*) pour générer toutes les règles $Y \rightarrow (X \setminus Y)(Y \subset X)$. Afin de limiter l'extraction aux règles d'association valides RAs, seules celles qui possèdent une confiance supérieure ou égale au seuil minimum min_{conf} sont retenues. Cet ensemble de règles peut ainsi être formalisé comme suit :

$$RAs = \{Y \rightarrow (X \setminus Y)/(X \in IF) \wedge (Y \subset X) \wedge confiance(Y \rightarrow (X \setminus Y)) \geq min_{conf}\} \quad (1.3)$$

Algorithme 2: Génération des k-motifs candidats avec $Apriori-Gen$

Input : Ensemble F_{k-1} des $k-1$-motifs fréquents
Output : Ensemble C_k des k-motifs candidats

```
1  begin
2      Insert into C_k
3      Select X.item_1, X.item_2, ..., X.item_{k-2}, X.item_{k-1}, Y.item_{k-1}
4      From F_{k-1} X, F_{k-1} Y
5      Where X.item_1 = Y.item_1, ..., X.item_{k-2} = Y.item_{k-2}, X.item_{k-1} < Y.item_{k-1}
6      for chaque candidat C de C_k do
7          for chaque sous-ensemble S de C de taille (k-1) do
8              if S ∉ F_{k-1} then
9                  Supprimer C de C_k
10     Retourner C_k
```

Une règle d'association est alors une implication entre deux motifs fréquents $X, Y \in IF$. Un des problèmes d'extraction de ces règles est la complexité qui est exponentielle avec la taille des motifs fréquents, où $2^k - 2$ règles d'association peuvent être générées pour un ensemble X de k-motifs fréquents.

Si nous reprenons l'exemple du "*panier*" (*table 1.1*), nous pouvons dégager des motifs fréquents (*ayant un support minimal supérieur ou égal à* 30%) de taille supérieure ou égale à deux. Ces motifs sont les suivants : {*Lait, Banane*}, {*Lait, Café*}, {*Lait, Pizza*}, {*Banane, Café*}, {*Banane, Pizza*} et {*Lait, Banane, Café*}, qui sont également illustrés dans la *table 1.4*. À partir de ces motifs fréquents, plusieurs règles d'association peuvent être engendrées. En fixant par exemple une confiance minimale $min_{conf} = 70\%$, seules les règles R_3, R_4, R_{12}, R_{13}, R_{15} et R_{16} parmi ces règles candidates, coloriées en magenta dans la *table 1.4* sont retenues.

Ainsi, la génération de l'ensemble de règles d'association passe par l'utilisation de mesures d'évaluation de celles-ci, telles que le *support* et la *confiance*. Certes, l'utilisation de ces deux mesures a été critiquée dans la littérature [AY98], [Fre99], [BVW03]. Nous trouvons par exemple que les algorithmes de type Apriori (*section 1.4.3*), c'est-à-dire basés sur ces deux mesures conduisent à l'obtention d'un nombre très important de règles, dont la plupart sont redondantes. En outre, il n'est pas aisé pour l'utilisateur de choisir les bonnes valeurs pour les seuils minimaux du support et de la confiance. De façon optimale, ces valeurs doivent dépendre de la taille et de la densité des données, mais les ressources informatiques sont limitées. Afin de pallier ce problème, plusieurs études ont été réalisées [PS91a], [AIS93], [KMR+94], [PT98a], [JA99], [PBTL99]. L'une des solutions proposées est l'utilisation de nouvelles mesures d'intérêt pour diminuer le nombre de règles [XL07], dont l'indicateur de Piatetsky-Shapiro [PS91a], la mesure

F_k	RAs	$Prémisse \rightarrow Conclusion$	$Support$	$Confiance$
$\{Lait, Banane\}$	R_1	$Lait \rightarrow Banane$	40%	50%
	R_2	$Banane \rightarrow Lait$	40%	67%
$\{Lait, Café\}$	R_3	$Lait \rightarrow Café$	60%	75%
	R_4	$Café \rightarrow Lait$	60%	100%
$\{Lait, Pizza\}$	R_5	$Lait \rightarrow Pizza$	40%	50%
	R_6	$Pizza \rightarrow Lait$	40%	67%
$\{Banane, Café\}$	R_7	$Banane \rightarrow Café$	40%	67%
	R_8	$Café \rightarrow Banane$	40%	67%
$\{Banane, Pizza\}$	R_9	$Banane \rightarrow Pizza$	40%	67%
	R_{10}	$Pizza \rightarrow Banane$	40%	67%
$\{Lait, Banane, Café\}$	R_{11}	$Lait, Café \rightarrow Banane$	40%	67%
	R_{12}	$Lait, Banane \rightarrow Café$	40%	100%
	R_{13}	$Café, Banane \rightarrow Lait$	40%	100%
	R_{14}	$Café, Lait \rightarrow Banane$	40%	67%
	R_{15}	$Banane, Lait \rightarrow Café$	40%	100%
	R_{16}	$Banane, Café \rightarrow Lait$	40%	100%

TABLE *1.4: Règles d'association extraites des données "paniers".*

intérêt [BMS97], l'indice de Gini [TKS04], etc.

1.5 Évaluation des règles d'association

Les mesures de règles sont utilisées comme étant un premier filtre pour apporter une solution au problème de l'extraction de règles. Face aux limites du support et de la confiance, de nombreuses mesures d'intérêt de règles d'association ont été présentées dans la littérature. Des mesures qui viennent compléter le *support* et la *confiance* et dont le but est de satisfaire les principaux objectifs de l'ECD, à savoir la production de règles intéressantes, facilement interprétables par l'utilisateur final. Cet utilisateur, qui est l'expert du domaine, est censé les valider pour choisir par la suite les meilleures règles [PS91a].

Les mesures d'intérêt [PS91a], [AS94], [KS96], [FPsS96], [BMS97], [AY98], [JA99], [Fre99], [LFZ99], [Zha00], [HLSL00], [BPT+00], [JS01], [HH01], [GKCG01], [Lal02], [AK02], [BH03], [LT04], [TKS04], [CFE05], [McG05], [TS06], [BLLV06], [GH07], [HGB+07], [LMVL08] jouent un rôle très important dans l'évaluation du degré d'intérêt et de pertinence des règles extraites [KMR+94], [ST96], [BMS97], [AY98], [LL99] en permettant un filtrage ou un ordonnancement automatique de celles-ci.

Il est aujourd'hui bien connu que l'évaluation objective d'une règle est particulièrement difficile sachant qu'une règle intéressante doit être "*valide, nouvelle et compréhensible*" [FPSSU96]. Les mesures d'intérêt peuvent être objectives (*ou mesures dirigées par les données*) ou subjec-

tives (*ou mesures dirigées par l'utilisateur*) [Fre99]. Idéalement, les mesures objectives peuvent être appliquées en premier lieu pour filtrer les règles non intéressantes et les mesures subjectives peuvent ensuite être appliquées pour identifier les règles qui sont d'un réel intérêt pour l'utilisateur.

1.5.1 Les mesures objectives

Une mesure objective prend en considération la structure des données et plus particulièrement les effectifs liés à la contingence des données [HH00], [TKS02], [LT04], [LMVL04] sans inclure aucune connaissance sur le domaine ni sur l'utilisateur [SG91], [Fre99]. Elle est basée essentiellement sur les données brutes des règles d'association extraites. La plupart des mesures objectives sont fondées principalement sur les théories de la statistique, probabilité, ou encore la théorie de l'information.

Définition 6 *(**mesure d'intérêt**) Une mesure d'intérêt est une fonction m de l'ensemble des règles d'association R à valeurs dans $\mathbb{R}$, telle que pour toute règle d'association $X \to Y$, $m(X \to Y)$ est calculée à partir des quatre quantités $n, n_X, n_Y, n_{X\overline{Y}}$.*

$$\textit{Une mesure d'intérêt est une fonction } m \;\left|\; \begin{array}{l} R \to \mathbb{R} \\ (X \to Y) \mapsto m(n, n_X, n_Y, n_{X\overline{Y}}) \end{array} \right.$$

Pour évaluer la qualité des connaissances découvertes, les mesures d'intérêt objectives sont utilisées dans différents domaines d'application de la fouille de données. McGarry et al. [MM04] ont utilisé ces mesures pour évaluer les règles extraites à partir de réseaux de neurones afin de découvrir leur fonctionnement interne. J. Azé [Azé03] a utilisé les mesures d'intérêt pour extraire des connaissances intéressantes et utiles pour l'expert à partir des données numériques et textuelles. W. Buntine [Bun96] a utilisé ces mesures pour explorer le modèle graphique probabiliste. Romao et al. [RFdSG04] ont utilisé les mesures d'intérêt dans un algorithme génétique qui optimise les croyances des experts pour classer les règles de prédiction floues selon leur intérêt. I. Kononenko [Kon95] a découvert les propriétés des mesures utilisées dans les arbres de décision. Également Gavrilov et al. [GAIM00] et Zhao et al. [ZK01] ont comparé les fonctions objectives utilisées dans les approches de clustering.

1.5.2 Les mesures subjectives

Une mesure subjective dépend principalement des objectifs, des connaissances et des croyances de l'utilisateur a priori sur le domaine étudié [ST95], [PT98b], [LHCM00]. En effet, elle permet de comparer les règles extraites par les mesures objectives avec les attentes de

l'utilisateur, ce qu'il veut et ce qu'il connait [LL99]. Les mesures subjectives permettent ainsi d'identifier le caractère inattendu et nouveau des règles découvertes par rapport aux croyances et au savoir de l'utilisateur.

Les critères subjectifs sont regroupés en quatre catégories :

- **Fiabilité** : Une règle est fiable si elle a une forte confiance, i.e., si la relation décrite par la règle se produit très souvent, presque dans la majorité des cas. Comme par exemple, la règle $(Acheter\ un\ ordinateur) \rightarrow (Acheter\ une\ imprimante)$.
- **Utilité** pour le problème posé : Une règle est utile si son contenu est riche d'information de qualité et que sa mise en pratique contribue à l'atteinte d'un objectif donné. Prenons par exemple la fameuse découverte de la grande chaine de distribution américaine Wal-Mart [3] : les gestionnaires de cette chaîne ont réalisés des tests et ont pu découvrir une règle qui leur a semblé importante, *l'achat groupé de couches pour bébés et de la bière le jeudi après midi* [LH96]. Cette découverte a encouragé Wal-Mart à entreprendre des actions par la réorganisation de ses rayons en mettant en place par exemple, le stock de couche de leur marque comme produit à grande marge bénéficiaire à côté du rayon des bières, ou encore de faire en sorte de vendre la bière et les couches pour bébés à prix entier le jeudi. Les actions effectuées par Wal-Mart en vue de cette révélation a pu augmenter considérablement leur chiffre d'affaire (*qui est leur principal objectif*). Afin d'exploiter ces actions, ces règles doivent trouver une explication [CYS03], [YH06], [LQ12].
- **Nouveauté** : La règle était inconnue. À titre d'exemple, une règle est nouvelle pour un utilisateur si ce dernier ne la connaissait pas avant et il est incapable de la déduire à partir d'autres règles connues [Sah99]. La nouveauté est ainsi détectée par l'utilisateur qui trouve que cette règle ne peut être déduite et ne contredit pas les règles précédemment découvertes.
- **Surprise** : La règle n'était pas attendue. Une règle est surprenante (*ou inattendue*) si elle contredit les connaissances ou les attentes de l'utilisateur [ST95], [LH96], [ST96]. Elle est aussi intéressante puisqu'elle permet d'identifier des défaillances dans des connaissances antérieures ainsi qu'elle peut suggérer un aspect des données qui nécessite une étude plus approfondie.

Les deux types de mesures (*objectives et subjectives*) décrites ci-dessus sont donc complé-

3. Wal-Mart charge quotidiennement des millions de transactions à partir de plus de 3600 magasins dans six pays et transmet en continu ces données pour son énorme entrepôt de données (*460 téraoctet*) Teradata (*Selon New York Times publié le 14 Novembre 2004.*)

mentaires, puisque les évaluations objectives des règles à l'aide de mesures d'intérêt, peuvent être complétées par des évaluations subjectives. En pratique, il est important d'utiliser les mesures objectives et subjectives en complémentarité pour extraire les connaissances intéressantes. Par exemple, deux règles d'association ayant les mêmes valeurs de confiance et de support n'auront pas la même utilité à un analyste qui cherche à résoudre un problème donné, d'où vient l'importance des critères de sélection subjectifs.

Dans cette thèse, nous allons nous concentrer uniquement sur les mesures objectives, qui seront analysées selon l'environnement d'étude décrit dans la section suivante.

1.5.3 Environnement d'étude

Face au problème de génération d'un nombre important de règles, l'une des solutions proposées étaient de les filtrer au moyen de mesures d'intérêt m. Pour ce faire, l'utilisateur doit fixer un seuil μ à chaque mesure et seules les règles R ayant une valeur supérieure à ce seuil sont retenues.

Cependant, la fixation du bon seuil pour l'ensemble de ces mesures s'avère problématique : avec un seuil μ trop bas, on risque d'avoir un nombre important de règles, et avec un seuil élevé, il serait possible qu'on élimine des règles intéressantes. Dès lors, diverses situations de référence sont proposées pour faciliter le choix des seuils. Pour chaque situation, il serait possible d'identifier les valeurs que peuvent prendre les mesures d'intérêt.

Par exemple, une valeur fixe prise par une mesure m dans l'une de ces situations, facilitera la tâche de l'utilisateur pour la détermination d'un seuil μ pour la mesure m vérifiant cette contrainte.

Les situations de référence : Nous présentons ci-dessous six situations de référence identifiées dans la littérature :

- *L'indépendance* : lorsque la réalisation de X n'affecte pas les chances de réalisation de Y, i.e., $P(XY) = P(X)P(Y)$ ou encore $P(Y|X) = P(Y)$;
- *L'équilibre ou indétermination* : lorsque X est réalisé, il y a autant de chances que Y ou $\overline{Y}$ soit réalisé, i.e., $P(Y|X) = 0.5$ (ou, $P(XY) = P(X)/2$) ;
- *L'implication logique* : lorsque Y est réalisé dès lors que X l'est, i.e., $P(Y|X) = 1$;
- *L'incompatibilité* : lorsque X et Y ne peuvent être réalisés simultanément, i.e., $P(Y|X) = 0$;
- *L'attraction* : lorsque la réalisation de X augmente les chances de réalisation de Y, i.e., lorsque $P(Y|X) > P(Y)$;
- *La répulsion* : lorsque la réalisation de X diminue les chances de réalisation de Y, i.e.,

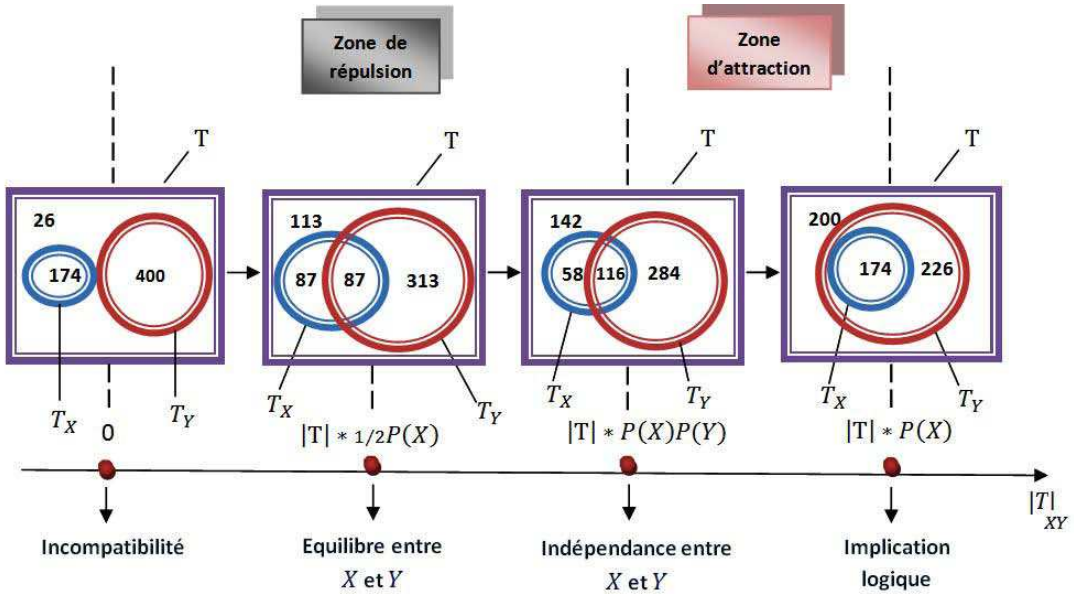

FIGURE *1.4: Les différents états d'une règle.*

lorsque $P(Y|X) < P(Y)$.

La *figure 1.4*, extraite de la thèse de S. Guillaume [Gui00], répertorie tous les états caractéristiques que l'on peut rencontrer, partant de l'incompatibilité jusqu'à l'implication logique. Deux zones (*de répulsion et d'attraction*) sont aussi identifiées, elles permettent d'exhiber le comportement qualitatif des règles d'association par rapport à l'indépendance. Les tailles de T_X, T_Y et T sont constantes et égales respectivement à 174, 400 et 600. Par contre, l'intersection entre T_X et T_Y peut varier. Nous nous appuyons sur cet environnement de travail pour étudier le comportement des mesures.

Seuls les trois dernières situations de référence mentionnées ci-dessus retiendront notre attention dans ce travail : l'indépendance, l'équilibre et l'implication logique. Nous écartons l'étude des mesures au niveau de l'incompatibilité puisqu'elle intéresse majoritairement les règles ayant une prémisse négative.

Dans la section suivante, nous exposons les différents travaux de la littérature réalisés sur les mesures d'intérêt objectives selon deux points de vue : tout d'abord selon une approche formelle, par la suite selon une approche empirique.

1.6 Étude formelle sur les mesures d'intérêt

Le comportement des mesures d'intérêt a suscité l'intérêt de plusieurs chercheurs. De multiples études ont alors été réalisées dans ce cadre : certains travaux se sont intéressés à la recherche de "*bonnes*" propriétés que devrait satisfaire une mesure d'intérêt ; tandis que d'autres

se sont basés sur ces propriétés pour catégoriser les mesures. Nous entamons cette section par une description des travaux sur les propriétés souhaitables d'une mesure.

1.6.1 Travaux sur les propriétés des mesures

Plusieurs travaux ont été menés dans le cadre de la recherche de "*bonnes*" propriétés pour les mesures objectives, dont ceux de [PS91a], que nous décrivons dans un premier lieu.

1.6.1.1 Travaux de Piatetsky-Shapiro [PS91a]

Piatetsky-Shapiro [PS91a] a proposé en 1991 trois propriétés (S_1 à S_3), que doit vérifier une mesure d'intérêt m :

- S_1 : la valeur de la mesure m doit être nulle dans le cas de l'indépendance ;
- S_2 : la mesure m doit être croissante en fonction du nombre n_{XY} d'exemples lorsque la taille n_X de la prémisse X et la taille n_Y de la conclusion Y restent constantes ;
- S_3 : la mesure m doit être décroissante en fonction de la taille n_X de la prémisse lorsque le nombre n_{XY} d'exemples et la taille n_Y de la conclusion restent constantes (*ou encore en fonction de la taille de la conclusion lorsque le nombre d'exemples et la taille de la prémisse restent constantes*).

De ces trois propriétés, il en déduit deux implications intéressantes qui sont les suivantes :

- I_1 : les valeurs des mesures doivent être positives en cas d'attraction entre X et Y ;
- I_2 : les valeurs des mesures doivent être négatives en cas de répulsion entre X et Y.

L'auteur a également proposé la mesure d'intérêt *Piatetsky-shapiro*, qui vérifie l'ensemble de ces propriétés (*voir sa définition en annexe A, mesure numéro 45*).

1.6.1.2 Travaux de Tan et al [TKS02]

Tan et al. ont réalisé en 2002 [TKS02] une étude sur 21 mesures (*Coefficient de corrélation, Confiance, Conviction, Cosinus, Facteur de certitude, Force collective, Gini, Goodman-Kruskal, Information mutuelle, Interêt, Jaccard, J-mesure, Laplace, Piatetsky-Shapiro, Q de Yule, Kappa, Klosgen, Ratio des chances, Support, Valeur ajoutée et Y de Yule*) et les ont évalué à travers huit propriétés. Les définitions de ces 21 mesures sont données dans l'*annexe A*.

Les trois premières propriétés se rapportent aux propriétés S_1 à S_3 de Piatetsky-Shapiro énoncées précédemment, et les cinq autres propriétés (T_{a1} *à* T_{a5}) sont les suivantes :

- T_{a1} : la mesure m est symétrique ;
- T_{a2} : la mesure m est invariante dans les deux cas suivants :
  - lorsqu'on multiplie les effectifs des ensembles T_{XY} et $T_{X\overline{Y}}$ par une constante positive k_1 et les effectifs des ensembles $T_{\overline{X}Y}$ et $T_{\overline{XY}}$ par une constante k_2 positive ;

- lorsqu'on multiplie les effectifs des ensembles $T_{X\overline{Y}}$ et $T_{\overline{XY}}$ par une constante positive k_1 et les effectifs des ensembles T_{XY} et $T_{\overline{X}Y}$ par une constante k_2 positive.

– T_{a3} : la mesure m doit vérifier les relations suivantes : $m(\overline{X} \to Y) = -m(X \to Y)$ et $m(X \to \overline{Y}) = -m(X \to Y)$;
– T_{a4} : la mesure m doit vérifier la relation $m(\overline{X} \to \overline{Y}) = m(X \to Y)$;
– T_{a5} : la mesure m doit être invariante lorsque la taille *n* de l'ensemble d'apprentissage T augmente et que tous les autres effectifs (n_X, n_Y *et* n_{XY}) restent constants.

Une étude comparative de ces propriétés selon 21 mesures d'intérêt a été effectuée, et a mis en exergue que chaque mesure possède des propriétés différentes la rendant utile pour certains domaines d'application, mais pas pour d'autres.

1.6.1.3 Travaux de Lallich et Teytaud [LT04]

Quant à Lallich et Teytaud dans [LT04], ils montrent les avantages et les inconvénients pour une mesure objective de posséder différentes propriétés. Cette discussion porte essentiellement sur 15 mesures (*Coefficient de corrélation, Confiance, Confiance centrée, Conviction, Facteur bayésien, Indice d'implication, Intensité d'implication, Intérêt, J-mesure, Loevinger, Moindre contradiction, Pearl, Piatetsky-Shapiro, Sebag-Schoenauer et Zhang*), mesures également répertoriées dans l'*annexe A*, et sur 13 propriétés qui sont les suivantes :

– L_1 : Intelligibilité ou compréhensibilité de la mesure [LMP+03], i.e.,
la mesure doit être intelligible pour pouvoir communiquer et expliquer les résultats obtenus.
– L_2 : Mesure non symétrique au sens de la négation de la conclusion, i.e.,
une mesure doit pouvoir faire la distinction entre $X \to Y$ et $X \to \overline{Y}$.
– L_3 : Mesure non symétrique, i.e.,
il est préférable d'avoir des mesures qui évaluent différemment les règles $X \to Y$ et $Y \to X$ puisque l'antécédent et la conclusion jouent des rôles différents [Fre99].
– L_4 : Mesure évaluant de la même façon $X \to Y$ et $\overline{X} \to \overline{Y}$ dans le cas de l'implication logique.
– L_5 : Taille de la prémisse fixe ou aléatoire, i.e.,
la taille de la prémisse est aléatoire lorsque la mesure est fondée sur un modèle probabiliste.
– L_6 : Mesure décroissante en fonction du nombre de contre-exemples, i.e.,
c'est une propriété équivalente à celle de Piatetsky-Shapiro, à savoir la propriété S_2.
– L_7 : Valeurs fixes pour l'indépendance (*généralisation de* S_1)
– L_8 : Valeurs fixes pour l'implication [LMP+03]
– L_9 : Tolérance aux premiers contre-exemples [GKCG01]

- L_{10} : Mesure croissante en fonction de la rareté du conséquent, i.e.,
  c'est une propriété équivalente à celle de Piatetsky-Shapiro, à savoir la propriété S_3.
- L_{11} : Mesure descriptive ou statistique, i.e.,
  une mesure est descriptive lorsque sa valeur est invariante en cas de dilatation des données, c'est-à-dire lorsque tous les effectifs sont multipliés par un même coefficient k. Dans le cas contraire, elle est statistique.
- L_{12} : Mesure discriminante, i.e.,
  c'est une mesure qui permet de discerner l'intérêt des règles, même lorsque l'ensemble d'apprentissage est volumineux. En effet, statistiquement une règle est d'autant plus fiable lorsqu'elle est évaluée sur un grand volume de données. Comme les mesures considèrent la taille de l'ensemble des données étudiées, elles deviennent peu discriminantes (*les règles sont jugées soit très bonnes avec des valeurs proches de 1, soit très mauvaises pour des valeurs proches de 0*) lorsque cette taille est grande.
- L_{13} : Facilité à fixer un seuil d'acceptation de la règle [LMPV03].

Les auteurs ont également apporté un éclairage intéressant sur les mesures et ont montré que plusieurs des mesures qu'ils ont étudiées peuvent s'exprimer comme des transformations affines de la confiance.

1.6.1.4 Travaux de Blanchard et al. [BGGB04]

Les auteurs de [BGGB04] préconisent que les mesures de qualité doivent tenir compte de l'état caractéristique "*équilibre*" ou "*indétermination*", c'est-à-dire l'état où le nombre d'exemples et de contre-exemples est identique. Pour cela, ils suggèrent que les mesures aient une valeur constante pour l'équilibre, d'où la propriété B_1 suivante :

- B_1 : valeur fixe dans le cas de l'équilibre.

Dans [BGBG5a], les auteurs ont également proposé une nouvelle mesure, l'indice probabiliste d'écart à l'équilibre (*voir annexe A, mesure numéro 26*), qui vérifie la propriété B_1 et qui calcule par conséquent la déviation de l'équilibre.

1.6.1.5 Travaux de Geng et Hamilton [GH07]

Geng et Hamilton dans [GH07] ont mené en 2007 une étude de 38 mesures (*Coefficient de corrélation, Confiance, Conviction, Cosinus, Couverture, Dépendance pondérée d'intérêt de Gray et Orlowska, Facteur bayésien, Facteur de certitude, Force collective, Gain informationnel, Gini, Goodman-Kruskal, Information mutuelle, Interêt, Jaccard, J-mesure, Klosgen, Laplace, Leverage, Loevinger, Moindre contradiction, Piatetsky-Shapiro, Précision, Prévalence, Q de Yule, Rappel, Ratio des chances, Risque relatif, Sebag-Schoenauer, Spécificité, Support, Support à*

sens unique, Support à double sens, Taux d'exemples et de contre-exemples, Valeur ajoutée, Variation du support à double sens, Y de Yule et Zhang) selon 11 propriétés. Toutes ces mesures sont rappelées en *annexe A* et l'ensemble des propriétés étudiées sont les suivantes. Tout d'abord, les propriétés S_1 à S_3 de Piatetsky-Shapiro puis les cinq propriétés T_{a1} à T_{a5} de Tan et al., et pour finir, les propriétés L_7 et L_9. La dernière propriété étudiée est la suivante :

– Le_1 : la mesure est croissante en fonction de la taille de l'ensemble d'apprentissage [Lal02], [LMVL04] et [GCB+04].

Une matrice d'évaluation de 38 mesures selon 11 propriétés est ainsi obtenue.

1.6.1.6 Travaux de Maddouri et Gammoudi [MG07]

Maddouri et Gammoudi ont étudié en 2007 dans [MG07] 62 mesures d'intérêt (*Confiance, Coefficient de corrélation, Support, Pavillon, Piatetsky-shapiro, Loevinger, Zhang, Indice d'implication, Intérêt, Moindre contradiction, Sebag-Schoenauer, Facteur bayésien, Conviction, Taux d'exemples, Kappa, Gain informationnel, Intensité d'implication, Intensité d'implication entropique, Indice probabiliste discriminant, Laplace, Confiance causale, Confirmation descriptive, Précision, Confiance-confirmée descriptive, Confiance-confirmée causale, Pearl, Cosinus, Ratio des chances, Q de Yule, Y de Yule, Jaccard, Klosgen, J-mesure, Gini, Rappel, Dépendance, Loevinger, Spécificité, Fiabilité négative, Kulczynski, Force collective, Indice probabiliste d'écart à l'équilibre, Dépendance causale putative, Ralambrodrainy, Indice de Wang, Ratio des liens, Interestingness, Ratio information mutuelle, Indice de Lerman, Erreur probabiliste de Chi2, Ratio d'information contraposée, Surprisingness, Gain, Ion, Gain rectangulaire, Indice de Rogers et Tanimoto, Indice Dice, Entropie de Shannon, Entropie quadratique, Ratio d'information, Quotient, Amélioration*) selon 12 propriétés. Parmi ces mesures, nous partageons les 43 premières mesures, que nous rappelons dans l'*annexe A*. Quant aux propriétés étudiées, les dix premières se rapportent aux propriétés : S_3, L_1, L_3, L_7, L_8, L_9, L_{10}, L_{13}, B_1, Le_1, et les deux propriétés restantes sont les suivantes :

– M_1 : Valeur fixe à l'incompatibilité
  La valeur de la mesure *m* doit être fixe dans le cas de l'incompatibilité, *i.e.,* lorsque $P(Y|X) = 0$;
– M_2 : Sensibilité d'une mesure au bruit [AK02]
  Idéalement, une mesure d'intérêt doit fournir des règles "stables", même si les données sont bruitées.

L'ensemble de ces travaux portent essentiellement sur les propriétés des mesures, où les chercheurs proposent des propriétés souhaitables pour une mesure d'intérêt permettant la compréhension de son comportement. Parmi ces études, certaines ne se sont pas limitées qu'à la

proposition de propriétés, tels que dans [PS91a], [LT04], [BGBG5a]. Ils ont aussi cherché à évaluer les mesures selon des propriétés qu'ils proposent ou qui existent déjà dans la littérature, comme dans les travaux de [TKS02], [GH07], [MG07] ou encore dans d'autres études s'intéressant à la classification des mesures d'intérêt.

1.6.2 Travaux sur la classification des mesures d'intérêt

La découverte de groupes de mesures d'intérêt ayant un comportement similaire est une étape importante, venant compléter les précédents travaux sur les "*bonnes*" propriétés des mesures.

1.6.2.1 Travaux de Blanchard et al. [BGBG5c]

Les auteurs dans [BGBG5c] se sont basés sur leurs travaux précédents [BGGB04], [BGBG5a] pour proposer une classification des mesures d'intérêt. Ayant défini dans [BGGB04] la propriété B_1, les auteurs catégorisent 19 mesures d'intérêt, dont les 15 premières mesures sont communes à nos deux travaux (*Confiance, Sebag, Taux d'exemples et de contre-exemples, Ganascia, Moindre contradiction, Coefficient de corrélation, Intérêt, Loevinger, Conviction, J-mesure, Ratio des chances, Facteur bayésien, Intensité d'implication, Indice d'implication, Indice de vraisemblance du lien, Contribution orientée à* X^2*, Règle d'intérêt, Indice d'inclusion, TIC*) selon les 2 critères suivants :

1. *le sujet* : la déviation de l'indépendance ou de l'équilibre [BGGB04] ;
2. *la nature* : descriptive ou statistique [LT04].

Étant donnés ces deux critères, les auteurs classifient les mesures d'intérêt en 4 catégories : (1) mesures descriptives/mesures de déviation de l'équilibre, (2) mesures descriptives/mesures de déviation de l'indépendance, (3) mesures statistiques/mesures de déviation de l'équilibre, et (4) mesures statistiques/mesures de déviation de l'indépendance. La majorité des mesures décrites étaient des mesures descriptives (L_{11}).

1.6.2.2 Travaux de Vaillant [Vai06]

Vaillant a étudié en 2006 dans [Vai06] 20 mesures d'intérêt (*Coefficient de corrélation, Cohen, Confiance, Confiance centrée, Conviction, Facteur bayésien, Indice d'implication, Indice probabiliste discriminant, Intensité d'implication, Intensité d'implication entropique tronquée, Intérêt, Gain informationnel, Laplace, Loevinger, Moindre contradiction, Piatetsky-Shapiro, Sebag-Schoenauer, Support, Taux d'exemples et de contre-exemples et Zhang*) selon 9 propriétés. Toutes ces mesures sont définies en *annexe A* et les propriétés étudiées sont les suivantes :

L_1, L_3, L_7, L_8, L_9, L_{10}, L_{13}, B_1 et Le_1. L'auteur affirme que ce sont les propriétés les plus intéressantes qu'une mesure doit vérifier. Il a alors étudié le comportement de ces 20 mesures par l'évaluation de celles-ci selon les 9 propriétés formelles citées. L'auteur a par la suite appliqué une méthode de classification hiérarchique (*CAH*) pour catégoriser les mesures et il a dégagé 5 classes de mesures.

Néanmoins, cette étude [LMVL08] ne s'intéresse qu'aux mesures qui décroissent en fonction de $n_{X\overline{Y}}$, reflétant par conséquent le biais des auteurs qui disent que moins il y a de contre-exemples, plus l'intérêt est élevé.

1.6.2.3 Travaux de Hyunh [Huy06]

Huynh a étudié en 2006 dans [Huy06] 36 mesures d'intérêt, dont 32 mesures communes à nos deux travaux (*Confiance, Laplace, Sebag-Schoenauer, Taux d'exemples, Confirmation descriptive, Confiance confirmée descriptive, Moindre contradiction, IPEE, Coefficient de corrélation, Intérêt, Loevinger, Conviction, Dépendance, Pavillon, J-mesure, Gini, Force collective, Ratio des chances, Q de Yule, Y de Yule, Klosgen, Kappa, Intensité d'implication, IIE, Support, Précision, Jaccard, Cosinus, Confiance causale, Confirmation causale, Confiance confirmée causale et Dépendance causale*) selon les 5 propriétés suivantes : B_1, L_7, L_{11}, T_{a1} et S_2. Cette étude a abouti à une matrice d'évaluation de 36 mesures d'intérêt selon 5 propriétés. En se basant sur cette matrice, l'auteur a catégorisé les mesures selon la "nature" et le "sujet", comme dans [BGBG5c], mais il a obtenu 5 groupes de mesures (*ce qui n'est pas le cas pour [BGBG5c] qui a trouvé 4 groupes seulement*). Pour 17 mesures supplémentaires étudiées par l'auteur [Huy06], en comparaison aux travaux de [BGBG5c], un nouveau groupe a été révélé qui comprend les mesures ne vérifiant aucun des deux critères (*nature et sujet*).

L'auteur a ensuite mis en exergue des relations mathématiques entre les mesures afin de découvrir des liens intéressants capables de limiter le nombre des mesures proposées. Certes, ces relations n'ont pas été considérées plus tard dans son travail.

1.6.2.4 Travaux de Feno [Fen07]

Feno en 2007 dans [Fen07] a étudié 15 mesures d'intérêt (*Coefficient de corrélation, Confiance, Confiance centrée, Conviction, J-mesure, Indice d'implication, Intérêt, Loevinger, Moindre contradiction, Nouveauté, Pearl, Piatetsky-Shapiro, Rappel, Sebag-Schoenauer et Support*) selon 13 propriétés. Les propriétés étudiées sont les suivantes : S_1, S_2, S_3, I_1, I_2, L_1, L_2, L_3, L_4, L_9, L_{13}, Le_1 et B_1. Cette même étude comportementale des mesures est présentée dans [TKS02]. Dans [Fen07], l'auteur s'est intéressé plus particulièrement à la mesure M_{GK} de [Gui00] et a proposé une approche analytique permettant de normaliser les 15 me-

sures selon M_{GK}. Il a obtenu 3 catégories de mesures : (i) mesures M_{GK} normalisables, (ii) mesures normalisables à normalisées différentes de M_{GK}, et (iii) mesures non normalisables.

1.6.2.5 Travaux de Heravi et Zaiane [HZ10]

Heravi et Zaiane ont mené une étude en 2010 dans [HZ10] de 53 mesures d'intérêt objectives, dont 45 mesures communes à nos deux travaux (*Support sens unique, Support double sens, Variation du support à double sens de Yao et Liu, Pavillon ou Confiance centrée, Facteur de certitude ou Loevinger, Force collective, Confiance, Confiance causale, Confirmation causale, Confirmation descriptive, Confiance-confirmée causale, Confiance-confirmée descriptive, Conviction, Coefficient de corrélation, Cosinus, Taux d'exemples, F-mesure ou Czekanowski-dice, Ganascia, Gini, Goodman-Kruskal, Indice d'implication, Gain informationnel, Intensité d'implication, Dépendance pondérée, Jaccard, J-measure, Kappa, Klosgen, Laplace, Moindre contradiction, Leverage, Intérêt, Loevinger, Information mutuelle, Multiplicateur de côte, Ratio des chances, Piatetsky-shapiro, Rappel, Risque relatif, Sebag-Schoenauer, Spécificité, Support, Q de Yule, Y de Yule, Zhang*) selon 16 propriétés. Nous identifions à travers cette étude uniquement les 11 propriétés suivantes : S_1, S_2, S_3, T_{a1}, T_{a2}, T_{a3}, T_{a4}, T_{a5}, L_7, L_9, Le_1. Les auteurs appliquent la méthode de classification hiérarchique sur la matrice d'évaluation de 53 selon 16 propriétés pour découvrir les mesures au comportement similaire.

1.6.2.6 Travaux de Le Bras [Bra11]

Le Bras [Bra11], [BLL12] a étudié 42 mesures (*Confiance, Confiance centrée, Conviction, Couverture, Moindre contradiction, Cosinus, Czekanowski-dice, Force collective, Gain informationnel, Gini, Intérêt, Jaccard, Kappa, Kulczynski, Nouveauté, Loevinger, Support sens unique, Support double sens, Piatetsky-shapiro, Prévalence, Rappel, Sebag, Spécificité, Taux d'exemples, Y de Yule, Facteur bayésien, Pearl, Gain, Ganascia, Indice d'implication, J-mesure, Klosgen, Laplace, Ratio des chances, Coefficient de pearson, Précision, Q de Yule, Risque relatif, Spécificité relative, Support, Valeur ajoutée et Zhang*) selon 6 propriétés opérationnelles qu'il a proposées. Au vu de ces 42 mesures, nous notons au total 38 mesures communes, dont certaines d'entre elles possèdent une même définition mais avec des noms différents [4].

Seules les mesures communes sont définies dans l'*annexe A*. L'auteur évalue les 42 mesures selon les 6 propriétés suivantes :

– Br_1 : Mesure robuste, i.e.,
il s'agit de tester la résistance de la mesure d'une règle par rapport à des perturbations de la base de données :

4. Intérêt représente la mesure *Pearl* dans nos travaux, *Levier* représente la mesure *Nouveauté* et *J1-mesure* correspond à la mesure *Support double sens*

- $Br_{1.1}$: Mesure plane, i.e.,
  pour certaines mesures, le calcul de la distance se ramène à un calcul de distance à un plan ce qui permet de fournir une solution algébrique exacte ;
- $Br_{1.2}$: Mesure quadratique, i.e.,
  les mesures élevées au carré, qui suivent une loi du carré. Une fonction quadratique s'écrit $y = f(x^2)$.

– Br_2 : Mesure efficace, i.e.,
  une mesure est efficace si elle permet de vérifier les propriétés algorithmiques suivantes :
  - $Br_{2.1}$: GUEUC, i.e.,
    c'est la propriété générale UEUC (*Universal Existential Upward Closure*) qui permet un élagage par monotonie descendante sur les règles ;
  - $Br_{2.2}$: Omni-monotonie, i.e.,
    cette propriété permet un élagage par anti-monotonie sur l'ensemble des motifs (*un motif est intéressant si toutes les règles qu'il engendre sont intéressantes*) ;
  - $Br_{2.3}$: Opti-monotonie, i.e.,
    la mesure permet d'éliminer la redondance ;

– Br_3 : Mesure anti-monotone.

Pour chacune des propriétés algorithmiques, l'auteur a fourni une généralisation de celles-ci (*GUEUC*, *omni-monotonie* et *opti-monotonie*) et il a proposé des conditions d'existence de ces généralisations.

Après avoir exposé la liste des propriétés (*ou critères*) permettant d'apprécier la qualité d'une mesure d'intérêt et présenté les différents travaux réalisés sur les mesures selon une approche formelle, nous illustrons dans ce qui suit un tableau synthétisant l'ensemble de ces travaux.

1.6.3 Tableau de synthèse de l'étude formelle

Dans cette sous-section, nous proposons un tableau de synthèse récapitulant les différentes études qui sont réalisées sur les mesures d'intérêt d'un point de vue théorique. Ce tableau de synthèse est illustré dans la *table 1.5*, où différents critères sont considérés. Dans ce tableau, nous exposons pour chaque étude formelle réalisée, le nombre de mesures analysées et qui varie entre 1 et 62 mesures, le nombre de propriétés nécessaires pour leur évaluation qui varie entre 3 et 13 propriétés. Nous indiquons ensuite la nature de ces propriétés qui peuvent porter sur la nature des mesures (*e.g., mesures statistiques, symétriques, etc.*) ou encore sur la formalisation des matrices (*e.g., invariance de la mesure suite à la permutation des lignes/colonnes*). Nous évoquons après la nature des études, nous trouvons que certains auteurs se sont inté-

ressés uniquement à la proposition des propriétés, tandis que d'autres ont cherché à évaluer les mesures selon les propriétés et à les catégoriser. Des techniques sont alors utilisées pour la classification des mesures comme par exemple la méthode de classification ascendante hiérarchique. Nous donnons finalement les résultats obtenus par chaque étude réalisée, tel que le nombre de groupes de mesures identifiés.

Dans ce qui suit, nous décrivons les études effectuées sur les mesures selon une approche empirique.

1.7 Étude empirique sur les mesures d'intérêt

Plusieurs études se sont intéressées à la comparaison des mesures d'intérêt d'un point de vue analyse de données. Nous commençons par présenter les travaux réalisés sur la classification des mesures.

1.7.1 Travaux sur la classification des mesures

Nous abordons dans ce qui suit les études portant sur la classification des mesures d'intérêt qui se basent sur les données.

1.7.1.1 Travaux de Hyunh et al. [HGB05a]

Huynh et al. [HGB05a] proposent une nouvelle approche basée sur l'analyse du graphe de corrélation, dont le but est de sélectionner des groupes de mesures fortement corrélées. Cette analyse est réalisée au moyen de l'outil ARQAT [HGB06], qui permet à l'utilisateur d'évaluer et comparer visuellement le comportement des différentes mesures. Les auteurs ont étudié dans un premier temps 34 mesures d'intérêt objectives et ont pu identifier onze groupes de mesures. Cette étude s'est basée sur les valeurs des mesures pour $120,000$ règles d'association découvertes à partir de la base de données *Mushroom*. Cette première présentation des résultats effectuée sur un seul jeu de données, a poussé les auteurs à enrichir leurs travaux [HGB$^+$07] par l'étude des mesures sur deux autres jeux de données de nature différente : une base fortement corrélée (*dense*) et une base faiblement corrélée (*éparse*). Ils se basent sur l'analyse de graphe de corrélation pour refléter le post-traitement des règles d'association et présenter les 5 groupes de mesures qu'ils ont identifiés.

Dans cette étude, les auteurs ont veillé à ce que les jeux de données sur lesquels ils ont travaillé aient des caractéristiques différentes. Néanmoins, leur nombre (2 *jeux de données*) remet en cause la robustesse et la validité de leurs résultats.

1.7.1.2 Travaux de Vaillant et al. [VLL04]

Lors de la mise en oeuvre d'un processus d'aide multicritère à la décision, Vaillant et al. [VLL04], [LVML07] ont effectué une catégorisation de 20 mesures en se basant à la fois sur des propriétés théoriques (*décrites dans la section 1.6.2.2 page 30*), et sur des résultats empiriques à partir de bases de données utilisant des matrices de comparaison de préordres. Cette étude expérimentale est réalisée sur 20 mesures selon 10 bases réelles provenant du répertoire UCI [5]. Pour étudier le comportement de ces mesures, les auteurs ont eu recours au développement d'une plateforme nommée "Herbs" [Vai02], permettant d'expérimenter les mesures sur des bases de règles. Cette étude consiste en effet à mesurer l'influence du choix de la mesure sur l'ordonnancement des règles en comparant les préordres engendrés par les mesures. Cette comparaison des préordres a permis de dégager 5 familles principales de mesures qui classent les règles de la même façon. Ils affirment également que certaines mesures sont des "transformées monotones croissantes de la confiance", tandis que d'autres sont des "transformées monotones croissantes de l'intérêt (*lift*)" ou encore "des transformées monotones entre-elles". Ces mesures seront capable de classer les règles selon un même ordre.

Les auteurs ont aussi comparé les résultats qu'ils ont obtenus par les deux approches (*formelle et empirique*) et concluent que les deux classifications sont très proches. Ils appliquent ensuite une approche d'aide multicritères à la décision [LMPV03] sur 8 propriétés de mesures afin d'aider l'utilisateur à choisir la/les mesure(s) d'intérêt la/les mieux adaptée(s) à ses besoins.

1.7.1.3 Travaux de Plasse et al. [PKSL06]

Plasse et al. [PKSL06] proposent une comparaison graphique de 5 mesures objectives pour évaluer l'intérêt des règles d'association. Pour ce faire, ils se sont appuyés sur l'étude menée par Lenca et al. [LMVL04] afin de sélectionner 3 mesures, auxquelles ils ont rajouté *Jaccard* et *Kulczynski*. Les auteurs ont effectué des expérimentations sur une base de données dense (*80 000 véhicules décrits par plus de 3 000 attributs binaires où chaque véhicule est décrit par la présence ou l'absence d'attributs binaires*) pour classer les règles selon les mesures d'intérêt. Ils déduisent que parmi les 5 mesures utilisées, *Jaccard* et *Kulczynski* sont celles qui permettent une meilleure sélection des règles issues de leur données. Deux groupes de mesures sont obtenus conduisant à des classements de règles identiques : (i) *Loevinger*, *Pavillon*, *Facteur bayésien* ; (ii) *Jaccard* et *Kulczynski*.

5. http ://ftp.ics.uci.edu/pub/machine-learning-databases

1.7.2 Travaux sur la génération/classement des règles

Nous abordons dans ce qui suit des études dressant essentiellement le problème de génération ou également de classement de règles.

1.7.2.1 Travaux de Tan et al [TKS02]

Tan et al. dans [TKS02] ont effectué dans un premier temps une étude comparative basée sur des propriétés de mesures, comme nous l'avons décrit dans la *section 1.6.1.2, page 26*. Dans cette même étude, les auteurs cherchent à trouver la mesure la mieux appropriée en utilisant des petits ensembles de tableaux de contingence. Pour ce faire, ils ont réalisé des expérimentations pour le classement de $10,000$ tables de contingence synthétiques générées par 21 mesures sur 6 jeux de données synthétiques et réels. Les auteurs ont réussi, suite à une variation des valeurs du support, à déterminer les paires de mesures corrélées. Afin de s'assurer des résultats obtenus empiriquement, Tan et al. se sont adressés à des experts dans les domaines choisis pour qu'ils ordonnent les tables de contingence dérivées des données, dans le but de choisir une mesure souhaitable. En confrontant les deux résultats, les auteurs ont conclu qu'aucune mesure n'est meilleure que les autres pour tous les domaines d'application. Ils pensent que différentes mesures ont des propriétés intrinsèques différentes, dont certaines pourraient être souhaitables pour certaines applications mais pas pour d'autres.

1.7.2.2 Travaux de Heravi et Zaiane [HZ10]

Heravi et Zaiane [HZ10] focalisent leur travail sur 53 mesures d'intérêt objectives pour la classification de règles associatives. Ainsi, une série de 20 bases de données prises du répertoire UCI sont utilisées pour étudier l'impact des mesures d'intérêt sur les classifieurs associatifs, qui représentent un sous-ensemble de règles d'association extraites (*i.e., ils calculent uniquement les règles ayant une valeur cible au niveau du conséquent*). Leur objectif porte sur l'effet du choix de mesures d'intérêt dans chacune de ces phases : génération/élagage de règles et sélection de règles, où ils cherchent à identifier la meilleure mesure pour chaque phase. Ils ont tout d'abord effectué une classification des mesures selon des propriétés, comme nous l'avons expliqué dans la *section 1.6.2.5, page 32*, et ils ont obtenu des groupes de mesures qui n'étaient que accessoires par rapport à leurs résultats, et qui n'ont pas vraiment été explorés plus loin dans leur papier.

Les auteurs concluent qu'il n'existe pas parmi les mesures étudiées, une qui a toujours donné les meilleurs résultats sur tous les ensembles de règles, et pour tous les jeux de données testés.

1.7.2.3 Travaux de Suzuki [Suz09b], [Suz09a]

En 2009, Suzuki [Suz09b], [Suz09a] s'est intéressé au domaine d'extraction de règles d'association. Dans [Suz09b], son intérêt a porté particulièrement sur les règles de classification en se basant sur le principe de longueur de description minimum (MDLP) pour la compression de données. Pour ce faire, l'auteur propose une mesure subjective basée sur MDLP, qu'il nomme "Longueur de codage négatif" intégrant les connaissances de l'utilisateur. Cette mesure pratique une recherche heuristique et construit une méthode de découverte de règles CLARDEM [6]. Des expérimentations approfondies sont réalisées au moyen de 10 bases Benchmark qui sont à la fois réelles et artificielles. L'auteur montre que sa méthode est capable de découvrir des groupes de règles de classification même à partir d'un petit ensemble de données bruitées.

Dans [Suz09a], Suzuki va au delà des règles de classification et s'intéresse aux règles d'association en général. Ainsi, il continue son travail précédent et s'appuie sur l'approche basée sur la compression des données pour présenter 2 exemples de mesures d'intérêt : la première mesure est la J-mesure [SG91] et la deuxième est une mesure basée sur une extension de la MDLP de CLARDEM, détaillée dans [Suz09b]. Ces deux mesures vérifient des propriétés intéressantes leur permettant d'éviter des pièges que l'auteur a souligné dans son article [Suz08], qui sont l'absence de paramètres et la robustesse contre le bruit [AK02]. Selon Suzuki, la J-mesure représente l'ensemble des mesures basées sur la compression, puisqu'elle décrit la quantité d'information compressée par une règle donnée. Tandis que la mesure basée sur une extension de la MDLP est une mesure subjective pour des groupes de règles de classification.

1.7.2.4 Travaux de Hébert et Crémilleux [HC06], [HC07]

Hébert et Crémilleux ont étudié en 2006 [HC06] le comportement des mesures d'intérêt usuelles d'extraction de règles d'association, mais ils se sont concentrés uniquement sur les règles de classification *i.e.* les règles dont la conclusion est une classe étiquetée. Pour ce faire, les auteurs proposent un environnement unificateur pour 17 mesures d'intérêt et prouvent que la majorité des mesures se comportent de la même manière. En termes de règles de classification, ils montrent que chaque mesure admet une borne inférieure à condition que deux paramètres soient introduits afin de caractériser au mieux la qualité d'une règle. Les auteurs indiquent aussi que toutes les mesures appartenant à cet environnement peuvent être simultanément optimisées, permettant alors la production des meilleures règles. Ils proposent une méthode pour extraire une couverture de règles optimisant les mesures, qui sont les règles de classification informatives. Des expérimentations sont alors réalisées sur un seul jeu de données (*mushroom*) afin de comparer les mesures d'intérêt et de mettre en valeur leurs similarités

6. Abréviation en anglais de "Classification Rule Discovery method based on an Extended-Mdlp"

et leurs différences, et aussi afin de quantifier la qualité des règles extraites en pratique par l'ensemble des 17 mesures.

En 2007, Hébert et Crémilleux [HC07] ont proposé une généralisation de leur précédent travail et ont réalisé une étude analytique des mesures d'intérêt objectives d'extraction de règles d'association. À travers cette étude, les auteurs cherchent alors à analyser le comportement des mesures d'intérêt et à montrer leurs caractéristiques communes. Ils expriment les mesures en fonction des fréquences dans le but de capturer leur effet de jointure puis ils exhibent les propriétés minimales (*2 propriétés de [PS91a] et une proposée par les auteurs*) qu'une mesure doit satisfaire afin d'avoir une vue unifiée de plusieurs mesures d'intérêt objectives, les SBMs (*Simultaneously lower Bounded Measure*). Un nouvel algorithme est proposé par les auteurs afin d'extraire les règles optimisant simultanément toutes les mesures de l'environnement. Ces règles sont des règles SBMs informatives, ayant des antécédents minimaux et des conséquents maximaux. Cette étude est réalisée sur un seul jeu de données réel du domaine médical et sur 17 mesures d'intérêt, où plusieurs paramètres doivent êtres introduits par l'utilisateur, tel que le nombre maximal d'exceptions de la règle. Selon les auteurs, le choix d'une mesure d'intérêt n'est pas vraiment posé puisque toutes les mesures se comportent semblablement.

Plusieurs travaux réalisés sur les mesures d'intérêt d'extraction de règles d'association sur les données ne considèrent pas l'avis d'un expert du domaine pour juger la qualité de l'information retenue, tels que les travaux cités précédemment. Cependant, nous nous apercevons que dans la littérature, de nombreuses études tiennent compte de l'avis subjectif de l'expert.

1.7.2.5 Travaux de Ohsaki et al. [OSK+04]

Un travail intéressant a été mené par Ohsaki et al. [OSK+04], qui étudient l'efficacité d'environ 20 mesures intérêt selon un jeu de données réel médical portant sur l'hépatite. Les auteurs comparent les valeurs des mesures avec les valeurs subjectives identifiées par l'utilisateur expert du domaine médical pour chaque règle donnée, i.e., l'intérêt humain réel. Les résultats montrent que certaines mesures, *Rappel* et la mesure χ^2 [MFM+98], peuvent prédire des règles intéressantes et que leur combinaison sera utile pour favoriser l'interaction du système humain. Ainsi, les mesures qu'ils ont considérées n'ont pas été comparées entre elles, mais plutôt aux connaissances subjectives de l'humain sur un ensemble de règles spécifiques.

1.7.2.6 Travaux de Carvalho et al. [CFE05]

Carvalho et al. [CFE05] proposent une étude de 11 mesures d'intérêt objectives selon 8 ensembles de données du répertoire UCI afin d'évaluer les règles de classification. Leur travail

se résume autour d'une question centrale : "Comment les mesures d'intérêt objectives sont-elles efficaces, dans le sens d'être de bons estimateurs du degrés d'intérêt subjectif d'une règle pour l'utilisateur ?". Cette question a été examinée suite à des expériences approfondies sur les mesures choisies par les auteurs, qui ont suivi 5 étapes permettant de répondre à cette question. L'avis d'un expert pour chaque ensemble de données est réalisé afin d'évaluer subjectivement l'intérêt des règles découvertes. Une confrontation des résultats de l'ordonnancement des règles de classification par les mesures avec ceux donnés par l'expert est ensuite effectuée, où il s'agit d'évaluer la corrélation entre les deux résultats. Les auteurs concluent qu'il n'existe pas un "*gagnant*" clair parmi les mesures objectives, puisque les valeurs de la corrélation qui sont associées à chacune des mesures varient considérablement entre les 8 jeux de données.

1.7.2.7 Travaux de Surana et al. [SKR10]

Surana et al. [SKR10] ont étudié 19 mesures d'intérêt selon 2 bases éparses réelles, prises du répertoire FIMI pour l'extraction de règles d'association rares. Pour ce faire, les auteurs proposent tout d'abord les propriétés qui doivent être vérifiées par les mesures pour extraire les règles d'association rares. Par la suite, ils ont réalisé une étude empirique sur les 19 mesures en suivant 5 étapes essentielles leur permettant de découvrir la mesure la mieux appropriée pour sélectionner les règles rares. Au cours de ce processus, Surana et al. ont eu recours à l'avis subjectif d'un expert dans le domaine en lui montrant un ensemble de 13 tables de contingence, dont certaines contiennent des variables fréquentes et rares, qu'il va ordonner. L'ordre des tables de contingence proposé par l'utilisateur expert a été confronté à celui révélé par les mesures d'intérêt afin de déterminer la ou les mesure(s) d'intérêt qui seront sélectionnées pour l'identification des règles d'association rares. Les auteurs ont trouvé que l'ordre identifié par les 3 mesures symétriques, *Jaccard, All-confidence* et *Cosinus* d'une part et les 3 mesures asymétriques, *Information mutuelle, Facteur de certitude* et *Pavillon* d'autre part, est le plus similaire à celui proposé par l'expert. Néanmoins, ils concluent qu'aucune mesure "*unique*" n'est appropriée pour extraire des règles d'association rares pour tous les jeux de données.

Ayant effectué une revue de synthèse de plusieurs études empiriques réalisées sur les mesures d'intérêt, nous proposons dans ce qui suit un tableau illustratif.

1.7.2.8 Tableau de synthèse de l'étude empirique

Dans la *table 1.6*, nous synthétisons l'ensemble des travaux réalisés sur les mesures d'intérêt selon une approche empirique, où 5 principaux critères sont considérés. Nous commen-

çons par présenter le nombre de mesures d'intérêt étudiées, variant entre 1 et 53 mesures selon l'étude ; ensuite le nombre de jeux de données analysés, où dans la majorité des travaux un seul jeu de données est repéré, et puis nous donnons leur nature. Ces jeux de données peuvent être de type réel, synthétique, épars ou dense. Le type du travail effectué est aussi un critère essentiel que nous proposons également, puisque certains chercheurs se sont intéressés à la catégorisation des mesures, à la différence d'autres qui ont considéré la génération et l'ordonnancement des règles dans leurs travaux. Les techniques qu'ils ont appliquées sont aussi présentées dans la *table 1.6*, pour certains ils ont eu recours à des plateformes d'expérimentation, où des algorithmes d'extraction de règles sont introduits. Les résultats identifiés par ces études représentent notre dernier critère. Ces résultats diffèrent d'un travail à l'autre : certains chercheurs ont obtenu des groupes de mesures tandis que d'autres ont montré qu'aucune mesure n'est meilleure que les autres, etc.

La section suivante va s'intéresser aux limites de l'existant et aux motivations de nos contributions.

1.8 Limites de l'existant et motivations

Les études réalisées sur les mesures d'intérêt objectives portent principalement sur deux axes de recherche complémentaires et essentiels : (1) étude formelle suite à la définition d'un ensemble de propriétés de mesures qui conduisent à une bonne évaluation de celles-ci ; (2) étude comparative expérimentale du comportement des différentes mesures d'intérêt d'un point de vue empirique. Compte tenu des deux tables de synthèses *1.5* et *1.6* qui ont dressé respectivement un bilan des différents travaux réalisés sur les mesures de façon formelle et empirique, nous dégageons certaines limites.

D'un point de vue formel, nous remarquons que plusieurs études s'intéressent au comportement des mesures et essentiellement à la catégorisation de celles-ci. Pour analyser le comportement des mesures, les chercheurs [TKS02], [LT04], [BGBG5a], [Vai06], [Huy06], [GH07] considèrent un nombre réduit de propriétés qu'ils jugent les plus importantes parmi celles existantes dans la littérature, ainsi qu'un nombre limité de mesures à catégoriser. En outre, en regardant les techniques utilisées par ces différents travaux (*table 1.5*), pour la classification des mesures, nous ne retrouvons qu'une seule méthode de classification utilisée, la méthode de classification ascendante hiérarchique [Vai06], [HZ10]. Certains autres travaux catégorisent les mesures selon des critères choisis à priori par les chercheurs [BGBG5c], [Bra11].

D'un point de vue empirique, nous remarquons des faiblesses dans les travaux réalisés essentiellement sur la classification des mesures d'intérêt de règles d'association [HGB05a], [LVML07], [PKSL06]. Ces faiblesses résident d'une part au niveau de la taille et de la densité

des jeux de données utilisés [VLL04], [LVML07], et d'autre part au niveau du nombre des jeux de données [HGB05a], [HC07] et de mesures analysées [VLL04], [HGB05a].

Cependant, si certains chercheurs ne se sont focalisés qu'à la classification des mesures, d'autres se sont intéressés à la génération et à l'ordonnancement des règles en considérant l'avis de l'utilisateur expert afin d'évaluer si les mesures peuvent prédire le degré subjectif d'une règle [OSK+04], [CFE05], [SKR10]. Cependant, les résultats obtenus manquaient de précision dans le sens où ni le nombre des mesures ni celui des bases de données testées ne sont importants [CFE05] et que parfois un seul expert est questionné [OSK+04]. Dans les travaux de [HC06], [HC07], les auteurs étudient le comportement d'une dizaine de mesures en se basant sur 3 propriétés formelles minimales (*essentiellement sur celles proposées par [PS91a]*), limitant ainsi leur travail.

Face à ces lacunes, nous proposons une étude approfondie du comportement des mesures d'intérêt selon deux axes de recherche : formel et empirique. Dans l'étude formelle, nous travaillons sur un nombre élevé de propriétés et de mesures d'intérêt, présentant ainsi une étude approfondie des mesures. Cette étude est le point de départ pour réaliser une classification de celles-ci, où différentes techniques sont utilisées. En effet, plus le nombre de mesures à catégoriser est élevé, meilleurs sont les résultats de la classification. Également, plus les méthodes de classification appliquées sont variées, plus robustes sont les classes de mesures obtenues. Dès lors, dans notre étude formelle des mesures, nous prenons ces critères en considération afin de proposer des classes de mesures robustes.

Dans notre étude empirique, nous cherchons à valider les résultats de la classification obtenue dans l'étude formelle. Pour ce faire, nous varions la taille et la nature des jeux de données utilisées. Le but de notre travail est d'aider l'utilisateur dans le choix de la "*bonne*" mesure, qui convient le plus à ses attentes et ses exigences par une approche par catégorisation.

1.9 Conclusion

L'un des problèmes attractifs en ECD, est celui de l'extraction de règles d'association. Dans ce chapitre, nous avons présenté les principales étapes de l'algorithme Apriori qui permet la génération d'un nombre important de règles. Afin de découvrir seulement les règles significatives et intéressantes, plusieurs mesures d'intérêt sont proposées dans la littérature. Mais vu le nombre de ces mesures, un défi supplémentaire est à relever : qui est le choix d'un "bon" ensemble de mesures d'intérêt de règles d'association, capable d'aider l'utilisateur pour identifier de la connaissance utile répondant à ses attentes. Ainsi, plusieurs travaux de la littérature se sont intéressés à ce problème de sélection de mesures. Nous trouvons des études formelles portant sur la proposition de propriétés de mesures qui conduisent à l'évaluation des mesures.

Ces propriétés font l'objet de la première partie du *chapitre 2*, qui synthétise et formalise les propriétés souhaitées d'une mesure. Aussi des études empiriques qui viennent compléter les études formelles, pour analyser le comportement des mesures selon les données.

Nous proposons dans le chapitre suivant un cadre formel pour étudier le comportement des mesures d'intérêt par l'évaluation de celles-ci selon les propriétés.

Auteurs	Nbre mesures	Nbre prop	Nature des propriétés	Nature de l'étude	Techniques utilisées	Résultats
Piatetsky-shapiro [PS91a]	1	3	Propriétés sur la nature des mesures	Proposition de propriétés	–	3 nouvelles propriétés de mesures
Tan et al. [TKS02]	21	8	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Proposition de propriétés + évaluation des mesures selon les propriétés	–	Étude comparative des mesures selon les propriétés
Lallich et Teytaud [LT04]	15	13	Propriétés sur la nature des mesures	Proposition de propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Étude comportementale des mesures
Blanchard et al. [BGBG5a] [BGBG5c]	19	4	Propriétés sur la nature des mesures	Proposition de propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Une nouvelle propriété de mesures ainsi qu'une nouvelle mesure sont proposées et 4 catégories de mesures ont été identifiées
Geng et Hamilton [GH07]	38	11	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Proposition de propriétés + évaluation des mesures selon les propriétés	–	Étude synthétique sur les mesures d'intérêt
Maddouri et Gammoudi [MG07]	62	12	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés	–	Étude théorique sur les mesures d'intérêt
B. Vaillant [Vai06]	20	9	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés + catégorisation des mesures	Classification ascendante hiérarchique	Identification de 5 groupes de mesures
X.-H. Hyunh [Huy06]	36	5	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Identification de 5 groupes de mesures
D. Feno [Fen07]	15	13	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Étude des propriétés des mesures + catégorisation des mesures	Catégorisation selon M_{GK} normalisable	Identification de 3 catégories de mesures
Heravi et Zaiane [HZ10]	53	11	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Catégorisation des mesures	Classification ascendante hiérarchique	Identification de groupes de mesures
Y. Le Bras [Bra11], [BLL12]	42	6	Propriétés algorithmiques + propriétés sur la nature des mesures	Proposition de propriétés + évaluation des mesures selon les propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Identification de 6 catégories de mesures

TABLE *1.5: Tableau de synthèse sur l'ensemble des travaux réalisés sur les mesures d'intérêt selon une approche formelle.*

Auteurs	Nbre mesures	Nbre bases	Nature des bases	Type d'étude	Techniques utilisées	Résultats
Hyunh et al. [HGB05a]	34	1	Réelle/dense	Catégorisation des mesures d'intérêt	Plateforme d'expérimentation ARQAT + classification basée sur le graphe de corrélation positive	Identification de 11 groupes de mesures
Hyunh et al. [HGB05b]	35	1	Réelle/dense	Catégorisation des mesures d'intérêt	approche d'analyse de données basée sur la distance entre les mesures + méthode de classification hiérarchique CAH + méthode de partitionnement k-médoide	Identification de 16 groupes de mesures
Hyunh et al. [HGB+07]	36	2	Dense/éparse	Catégorisation des mesures d'intérêt	Plateforme d'expérimentation ARQAT + classification basée sur le graphe de corrélation	Identification de 5 groupes de mesures
Vaillant et al. [VLL04] [LVML07]	20	10	Réelle	Catégorisation des mesures d'intérêt	Plateforme d'expérimentation HERBS + comparaison de préordres	Identification de 5 groupes de mesures
Plasse et al. [PKSL06]	5	1	Réelle/dense	Classement des règles + catégorisation des mesures	Représentation graphique originale basée sur des courbes de niveaux	Identification de deux groupes de mesures
Tan et al. [TKS02]	21	6	1 base synthétique + 5 bases réelles	Calcul de la similarité entres les paires de mesures + classement des tables de contingence + avis d'un expert	Calcul de la corrélation au moyen de l'indice de Pearson + ordonnancement au moyen d'un algorithme qu'ils proposent	Aucune mesure n'est meilleure que les autres pour tous les domaines d'application
Heravi et Zaiane [HZ10]	53	20	Réelle	Classification de règles associatives	Classifieurs associatifs	Il n'existe pas de mesure "*unique*" qui a un impact sur tous les ensembles de règles pour tous les jeux de données
Suzuki [Suz09b]	1	10	Réelle/artificielle	Génération de règles de classification	Méthode de découverte de règles CLARDER qu'ils ont proposé	Découverte de groupes de règles de classification
Hébert et Crémilleux [HC06]	17	1	Réelle/dense	Étude comparative des mesures d'intérêt + génération de règles de classification informatives	Proposition d'un environnement unificateur des mesures d'intérêt + d'une méthode de découverte de règles de classification	La majorité des mesures se comportent de la même façon
Hébert et Crémilleux [HC07]	17	1	Réelle/dense	Étude comparative des mesures d'intérêt + génération de règles d'association	Un nouvel environnement + nouveau algorithme sont proposés afin d'extraire les règles optimisant l'ensemble des mesures	Le choix d'une mesure d'intérêt n'est pas vraiment posé puisque toutes les mesures se comportent semblablement
Ohsaki et al. [OSK+04]	20	1	Réelle	Génération de règles d'association	Système d'extraction fondé sur un cadre typique de fouille de données, séries chronologiques + avis de l'expert	La combinaison de certaines mesures favorise l'interaction du système humain
Carvalho et al. [CFE05]	11	8	Réelle	Génération et ordonnancement de règles de classification	Calcul de la corrélation + avis de l'expert	Il n'existe pas un "*gagnant*" clair parmi les mesures
Surana et al. [SKR10]	19	2	Réelle/éparse	Étude analytique pour sélectionner la bonne mesure d'extraction de règles d'association rares	Proposition d'un environnement d'analyse des mesures + recours à l'avis de l'expert	Aucune mesure unique n'est appropriée pour extraire les règles d'association rares pour tous les jeux de données

TABLE *1.6: Tableau de synthèse sur l'ensemble des travaux réalisés sur les mesures d'intérêt selon une approche empirique.*

Points clés

Positionnement :

– Revue de la littérature sur les différentes études réalisées sur les mesures d'intérêt objectives selon une approche théorique et empirique.

CHAPITRE 2

Étude des mesures d'intérêt

Sommaire

2.1 Introduction

Afin de guider l'utilisateur expert vers les connaissances potentiellement intéressantes, des mesures d'intérêt de règles d'association sont utilisées, dont l'objectif est d'évaluer la qualité des règles. Deux mesures sont classiquement utilisées, le *support* et la *confiance*, mais elles sont insuffisantes pour garantir la qualité des règles révélées. Elles ont été remises en cause dans différents travaux comme par exemple [SM02]. Afin de surmonter ce problème, de nombreuses mesures d'intérêt objectives ainsi que plusieurs propriétés souhaitables [PS91a], [GKCG01], [TKS02], [LMPV03], [LT04], [BGBG5a] pour concevoir ou évaluer une "*bonne*" mesure ont été proposées dans la littérature. Des études formelles ont ainsi été réalisées sur les mesures d'intérêt [TKS02], [LT04], [Vai06], [GH07], [Fen07], [HGB+07], [HZ10], [Bra11], *voir chapitre 1*. Ces études, portent sur un nombre restreint de mesures et de propriétés. Face à cette limite (*cf. table 1.5*), nous abordons à travers ce chapitre une étude approfondie d'un grand nombre de mesures selon plusieurs propriétés formelles. Pour ce faire, nous proposons une formalisation de ces propriétés afin de lever toute ambiguïté sur celles-ci, ainsi qu'une généralisation à chaque fois qu'une propriété l'autoriserait. Cela nous permet alors de construire un cadre formel afin d'étudier convenablement le comportement des différentes mesures.

La *section 2.2* de ce chapitre présente un recensement de quelques mesures d'intérêt parmi celles présentes dans la littérature. Une formalisation de l'ensemble des propriétés de mesures synthétisées dans le *chapitre 1* est effectuée dans la *section 2.3*, suivie d'une évaluation des mesures selon les propriétés mises en évidence précédemment. Nous terminons ce chapitre par l'identification de relations mathématiques entre les mesures et la proposition d'une matrice de mesures-propriétés, qui sera le point de départ pour une catégorisation des mesures d'intérêt.

2.2 Les mesures d'intérêt objectives

Nous présentons dans cette section l'ensemble des mesures d'intérêt ayant été utilisées dans les études réalisées dans le cadre de cette thèse. Nous donnons leur expression et nous interprétons sémantiquement quelques exemples de mesures.

2.2.1 Liste des mesures utilisées

Il existe dans la littérature une pléthore de mesures d'intérêt. Nous en avons recensé 69 mesures dont 46 proviennent des travaux de synthèse de [PS91a], [TKS02], [LT04], [Vai06], [GH07], [Fen07]. Neuf mesures ont été extraites de l'étude réalisée par [HGB+07], et sont les suivantes : *Confiance causale, Confiance confirmée causale, Confiance confirmée descriptive,*

Confirmation causale, Confirmation descriptive, Dépendance, Dépendance causale estimée, Pavillon et Support causal. Les mesures restantes proviennent de différents travaux et sont : *Czekanowski-Dice* [Cze13], *Fukuda* [FMMT96], *Ganascia* [Gan87], *Indice probabiliste d'écart à l'équilibre* [BGBG5a], *Indice probabiliste d'écart à l'équilibre entropique* [BGBG5b], *Intensité d'implication entropique* [GKCG01], *Indice de vraisemblance du lien* [Ler70], *Kappa* [Coh60], *Kulczynski* [Kul28], M_{GK} [Gui00], *Ochiai* [Och57], *Satisfaction* [LFZ99] et *VT100* [RM08].

À travers l'étude de ces différentes mesures, nous avons détecté des mesures identiques mais portant des noms différents. Ces mesures sont les suivantes : *Coefficient de corrélation ou ϕ-coefficient*, *Cohen ou Kappa*, *Confiance centrée ou Valeur ajoutée ou Pavillon*, *Confiance confirmée descriptive ou Ganascia*, *Cosinus ou Ochiai*, *Czekanowski-Dice ou F-mesure*, *Facteur bayésien ou Multiplicateur de côte*, *Facteur de certitude ou Loevinger ou Satisfaction*, *Kulczynski ou Indice d'accord et de désaccord*, *Support ou Indice de Russel et Rao* et *Précision ou Support causal*. En gardant une seule mesure par groupe de mesures sémantiquement identiques, nous nous retrouvons avec 61 mesures d'intérêt au lieu de 69.

Ces 61 mesures étudiées dans [GGM10], sont résumées dans les *tables 2.1, 2.2* et regroupées en deux catégories : les mesures symétriques, d'une part, et non d'autre part. Les expressions de chacune des mesures, accompagnées de leur référence, sont disponibles dans [GGM] et rappelées en *annexe A*. Nous ordonnons ces mesures par ordre alphabétique et afin de faciliter la recherche de leurs définitions dans l'*annexe A*, nous donnons leurs numéros dans les *tables 2.1 et 2.2*.

2.2.2 Interprétation de quelques exemples de mesures objectives

Nous donnons dans ce qui suit quelques exemples de mesures d'intérêt tout en essayant de les interpréter sémantiquement. Le choix de ces mesures était fait aléatoirement.

Confiance causale [Kod01] : La mesure *Confiance causale* (*ConfCaus*) est définie selon Y. Kodratoff par l'expression suivante :

$$ConfCaus(X \rightarrow Y) = 1 - \frac{1}{2}\left(\frac{1}{P(X)} + \frac{1}{P(\overline{Y})}\right) P(X\overline{Y}) = 1 - \frac{1}{2}\, P(\overline{Y}/X) - \frac{1}{2}\, P(X/\overline{Y})) \tag{2.1}$$

C'est l'addition de la confiance due aux instances directes de la règle $(X \rightarrow Y)$ et la confiance apportée par sa contraposée $(\overline{Y} \rightarrow \overline{X})$. Le coefficient $1/2$ introduit dans sa définition provient du fait que chaque probabilité (*a posteriori*) peut être égale à 1.

Mesures symétriques			
1	φ-coefficient *ou* Coefficient de corrélation	2	Cohen *ou* Kappa
11	Cosinus *ou* Ochiai	13	Czekanowski-Dice *ou* F-mesure
20	Force collective	22	Gain informationnel
24	Goodman	33	Indice de vraisemblance du lien (IVL)
34	Intérêt *ou* Lift	35	Jaccard
38	Kulczynski *ou* Indice d'accord désaccord	43	Nouveauté
44	Pearl	45	Piatetsky-Shapiro
46	Précision *ou* Support causal	48	Q de Yule
50	Ratio des chances	54	Support *ou* Indice de Russel et Rao
56	Support à double sens de Yao et Liu (SDS)	58	VT100
59	Variation support à double sens de Yao et Liu (VS)	60	Y de Yule

TABLE *2.1: Les mesures d'intérêt objectives symétriques étudiées.*

Confiance centrée [LT04] : La *Confiance centrée* (*ConfCent*) ou *Valeur ajoutée* ou *Pavillon* est exprimée comme suit :

$$ConfCent(X \rightarrow Y) = P(Y/X) - P(Y) = \frac{P(XY) - P(X)P(Y)}{P(X)} = P(\overline{Y}) - P(\overline{Y}/X) \quad (2.2)$$

Cette mesure est une transformation affine de la Confiance, proposée afin de remédier aux défauts de cette dernière, qui sont sa variabilité à l'indépendance et son insensibilité à la taille de l'ensemble d'apprentissage. Ainsi, pour changer le comportement de la confiance pour certaines propriétés, les auteurs [LT04] ont pensé à introduire $P(Y)$ à la mesure en faisant un recentrage de la confiance par rapport à la taille de la conclusion de la règle $X \rightarrow Y$.

Conviction [BMS97] : La *Conviction* est l'une des mesures qui favorisent les contre-exemples de la règle $X \rightarrow Y$. Elle est définie par :

$$Conviction(X \rightarrow Y) = \frac{P(X)P(\overline{Y})}{P(X\overline{Y})} \quad (2.3)$$

La *conviction* est utilisée pour quantifier l'écart à l'indépendance de la règle $X \rightarrow Y$. Elle peut être interprétée de cette façon : si la valeur de la conviction pour une règle donnée est égale à 1, alors les attributs X et Y sont indépendants. Si sa valeur vaut 2, ceci montre qu'il est

Mesures non symétriques			
3	Confiance *ou* Précision	4	Confiance causale
5	Confiance centrée *ou* Valeur ajoutée *ou* Support changé *ou* Pavillon	6	Confiance confirmée descriptive *ou* Ganascia
7	Confiance confirmée causale	8	Confirmation causale
9	Confirmation descriptive	10	Conviction
12	Couverture	14	Dépendance
15	Dépendance causale	16	Dépendance pondérée
17	Facteur bayésien *ou* Multiplicateur de côte	18	Facteur de certitude *ou* Satisfaction *ou* Loevinger
19	Fiabilité négative	21	Fukuda
23	Gini	25	Indice d'implication
26	Intensité probabiliste d'écart à l'équilibre (IPEE)	27	Intensité probabiliste entropique d'écart à l'équilibre (IP3E)
28	Indice probabiliste discriminant (IPD)	29	Information mutuelle
30	Intensité d'implication (II)	31	Intensité d'implication entropique (IIE)
32	Intensité d'implication entropique révisée (IIER)	36	J-mesure
37	Klosgen	39	Laplace
40	Leverage	41	M_{GK}
42	Moindre contradiction *ou* Surprise	47	Prévalence
49	Rappel	51	Risque relatif
52	Sebag-Schoenauer	53	Spécificité
55	Support à sens unique	57	Taux d'exemples
61	Zhang		

TABLE *2.2: Mesures d'intérêt objectives non symétriques étudiées.*

deux fois plus convaincant de considérer X et Y selon leur relation par la règle $X \rightarrow Y$. Ainsi, plus la valeur de la conviction est élevée, plus le nombre de contre-exemples diminue jusqu'à avoir des règles ne possédant que des exemples.

Facteur bayésien [Jef35] : Le *Facteur bayésien* étant une forme des ratios des chances (*odds ratio*). Il s'exprime de la manière suivante :

$$Facteur\ bayésien(X \rightarrow Y) = \frac{P(XY)P(\overline{Y})}{P(X\overline{Y})P(Y)} \tag{2.4}$$

C'est le rapport de deux probabilités conditionnelles inversées au niveau du nombre du conséquent.

Fiabilité négative [LFZ99] : La *Fiabilité négative* permet de mesurer la fiabilité de la règle à prédire les cas négatifs. Elle est définie de la manière suivante :

$$Fiabilité\ négative(X \rightarrow Y) = P(\overline{X}/\overline{Y}) \tag{2.5}$$

Cette mesure est semblable à la confiance qui permet de prédire les cas positifs de la règle.

Intérêt [BMS97] : Étant donnée la règle $X \rightarrow Y$, la mesure *Intérêt* ou *Lift* représente le rapport de la probabilité jointe de X et Y au produit des probabilités marginales. Elle est exprimée de la manière suivante :

$$Intérêt(X \rightarrow Y) = \frac{P(XY)}{P(X) \times P(Y)}. \tag{2.6}$$

L'intérêt permet ainsi de déterminer la distance de la règle à l'indépendance. Si la valeur de l'intérêt est de 1, alors X et Y sont indépendants. Cependant, plus cette valeur est élevée, plus il est probable que la présence de X et Y ensemble lors d'une transaction n'est pas seulement un événement aléatoire, mais c'est à cause d'une relation entre eux.

Indice d'implication [LGR81], [LA07] : *L'indice d'implication* (*IndImp*) est construit en suivant une loi de poisson ayant comme paramètres $n_X.P(\overline{Y})$. Il est défini de la manière suivante :

$$IndImp(X \rightarrow Y) = \sqrt{n}\ \frac{P(X\overline{Y}) - P(X)P(\overline{Y})}{\sqrt{P(X)P(\overline{Y})}} \tag{2.7}$$

Cet indice permet d'évaluer le nombre de contre-exemples à la règle $X \rightarrow Y$ par rapport au nombre attendu sous l'hypothèse de l'indépendance. En effet, plus la valeur de l'indice d'impli-

cation est élevée, plus le nombre de contre-exemples de la règle augmente par rapport à une situation d'indépendance.

Intensité d'implication [Gra79] : *L'intensité d'implication* (II) est une mesure probabiliste qui, semblablement à l'indice d'implication, permet d'évaluer le nombre de contre-exemples par rapport au nombre attendu sous l'hypothèse de l'indépendance. Gras et al. [GAB+96] ont construit cette mesure en suivant la loi de poisson de paramètres $n.P(X).P(\overline{Y})$ comme suit :

$$II(X \rightarrow Y) = P\Big[Poisson(nP(X)P(\overline{Y})) \geq P(X\overline{Y})\Big] \tag{2.8}$$

Il est possible de s'approcher de la loi normale centrée réduite en écrivant l'intensité d'implication à partir de l'indice d'implication comme suit :

$$II(X \rightarrow Y) = 1 - \phi(IndImp)] \tag{2.9}$$

où ϕ est la fonction de répartition de la loi normale centrée réduite.

Lorsque la taille des données est grande, l'intensité d'implication tend vers soit 0 soit 1. [Gui00].

Loevinger [Loe47] : L'indice de *Loevinger* ou *Facteur de certitude* ou *Satisfaction* est parmi les plus anciens indices répertoriés dans la littérature avec le support et la confiance. Il est défini par :

$$Loevinger(X \rightarrow Y) = \frac{P(Y/X) - P(Y)}{1 - P(Y)} = \frac{P(Y/X) - P(Y)}{P(\overline{Y})} = 1 - \frac{P(X\overline{Y})}{P(X)P(\overline{Y})} \tag{2.10}$$

Cet indice permet d'améliorer l'interprétation de la confiance d'une règle par une normalisation de celle-ci. Il s'agit de normaliser la confiance centrée de la règle par rapport aux exemples ne vérifiant pas le conséquent de la règle.

M_{GK} **[Gui00] :** M_{GK} est une mesure d'intérêt qui permet l'extraction de règles négatives. Elle s'exprime de la manière suivante :

$$M_{GK} = \frac{p(Y/X) - p(Y)}{1 - p(Y)}, \text{ si X favorise Y; } M_{GK} = \frac{p(Y/X) - p(Y)}{p(Y)}, \text{ autrement} \tag{2.11}$$

Cette mesure prend en considération plusieurs situations de référence : En cas où la règle est située dans la zone attractive, cette mesure évalue la distance entre l'indépendance et l'implication logique. Ainsi, plus la valeur de M_{GK} est proche de 1, plus la règle est proche de

l'implication logique ; et plus la valeur de M_{GK} est proche de 0, plus la règle est proche de l'indépendance. Dans le cas où la règle se trouve dans la zone répulsive, M_{GK} évalue cette fois la distance entre l'indépendance et l'incompatibilité. Ainsi, plus la valeur de M_{GK} est proche de -1, plus la règle est semblable à l'incompatibilité ; et plus la valeur de M_{GK} est proche de 0, plus la règle est proche de l'indépendance.

Sebag-Schoenauer [SS88] : Cette mesure prend en considération le nombre de contre-exemples de la règle $X \rightarrow Y$. Elle est définie de la manière suivante :

$$Sebag(X \rightarrow Y) = \frac{P(XY)}{P(X\overline{Y})} \tag{2.12}$$

C'est le rapport entre le nombre d'exemples et le nombre de contre-exemples de la règle. Cette mesure peut-être interprétée de la manière suivante : une règle ayant une valeur de Sebag inférieure à 1 possède plus de contre-exemples que d'exemples et inversement.

Support à sens unique de Yao et Liu (SU) [YZ99] : La mesure *Support à sens unique de Yao et Liu* utilise les quantités suivantes pour mesurer la signification de la règle $X \rightarrow Y$:

$$SU(X \rightarrow Y) = P(Y/X) log_2 \frac{P(XY)}{P(X)P(Y)} \tag{2.13}$$

Cette mesures est définie comme étant le produit de la *Confiance* et du logarithme de la mesure *Intérêt* qui permet de quantifier l'écart à l'indépendance de la règle $X \rightarrow Y$. Plus la valeur du support à sens unique est élevée, plus les attributs X et Y sont corrélés.

La section suivante portera sur les propriétés souhaitables d'une mesure d'intérêt, qui seront utilisées pour étudier les caractéristiques des diverses mesures définies précédemment.

2.3 Synthèse et formalisation des propriétés des mesures

Afin d'étudier le comportement de l'ensemble des mesures présentées en *annexe A*, nous retenons 22 propriétés formelles. Ces propriétés sont illustrées dans la *figure 2.1* et ont été synthétisées dans le *chapitre 1*.

Dans cette section, nous nous intéressons à l'interprétation et à la formalisation de ces 22 propriétés. Cette formalisation va apporter une nouveau regard sur ces propriétés, permettant ainsi d'enlever toute ambiguité sur celles-ci.

No.	Propriétés	Réf.
P_1	Intelligibilité ou compréhensibilité de la mesure	[LMPV03]
P_2	Facilité à fixer un seuil	[LT04]
P_3	Mesure non symétrique.	[TKS04], [LT04]
P_4	Mesure non symétrique dans le sens de la négation de la conclusion.	[LT04], [TKS04]
P_5	Mesure évaluant de la même façon $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ dans le cas de l'implication logique.	[LT04]
P_6	Mesure croissante en fonction du nombre d'exemples ou décroissante en fonction du nombre de contre-exemples.	[PS91a], [LT04]
P_7	Mesure croissante en fonction de la taille de l'ensemble des données.	[TKS04], [GH07]
P_8	Mesure décroissante en fonction de la taille de la conclusion ou de la taille de la prémisse.	[LT04], [PS91a]
P_9	Valeur fixe a dans le cas de l'indépendance.	[PS91a] , [LT04]
P_{10}	Valeur fixe b dans le cas de l'implication logique.	[LT04]
P_{11}	Valeur fixe c dans le cas de l'équilibre.	[BGBG5a]
P_{12}	Valeurs identifiables dans le cas de l'attraction entre X et Y.	[PS91a]
P_{13}	Valeurs identifiables dans le cas de répulsion entre X et Y.	[PS91a]
P_{14}	Tolérance aux premiers contre-exemples.	[LT04], [Vai06]
P_{15}	Invariance en cas de dilatations de certains effectifs.	[TKS04]
P_{16}	Opposition des règles $X \rightarrow Y$ et $\overline{X} \rightarrow Y$.	[TKS04]
P_{17}	Opposition des règles antinomiques $X \rightarrow Y$ et $X \rightarrow \overline{Y}$.	[TKS04]
P_{18}	Égalité entre les règles $X \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$.	[TKS04]
P_{19}	Mesure fondée sur un modèle probabiliste ou non.	[LT04]
P_{20}	Mesure descriptive ou statistique.	[LT04]
P_{21}	Mesure discriminante.	[LT04]
P_{22}	Mesure robuste.	[BMLL10a]

FIGURE *2.1: Les propriétés des mesures d'intérêt.*

2.3.1 Propriété 1 : Intelligibilité ou compréhensibilité de la mesure

Cette propriété est reprise de la propriété L_1 (*Intelligibilité ou compréhensibilité de la mesure*, [LT04]) et nous conservons les 3 niveaux d'intelligibilité définis par [LMP+03].

$$P_1(m) = 0 \quad si \quad l'interprétation\ de\ m\ est\ difficile,$$
$$P_1(m) = 1 \quad si \quad m\ se\ ramène\ à\ des\ quantités\ usuelles,$$
$$P_1(m) = 2 \quad si \quad m\ peut\ s'expliquer\ par\ une\ phrase.$$

Il est important qu'une mesure soit intelligible pour pouvoir interpréter les résultats obtenus. Le support et la confiance sont deux mesures ayant un sens "concret" puisqu'elles peuvent s'expliquer par une phrase simple, comme :

- Le support d'une règle $X \rightarrow Y$ représente "*la fréquence d'apparition de cette règle*",
- La confiance d'une règle $X \rightarrow Y$ représente "*la force de cette règle*".

Nombreuses sont les mesures d'intérêt qui ne vérifient pas cette propriété, rendant ainsi leurs valeurs difficilement interprétables par l'utilisateur.

2.3.2 Propriété 2 : Facilité à fixer un seuil d'acceptation de la règle

Cette propriété est reprise de la propriété L_{13} (*Facilité à fixer un seuil d'acceptation de la règle*, [LMPV03]). Elle est proposée dans le but de conserver les règles intéressantes sans

avoir à les classer. Les auteurs dans [Vai06] pensent que les mesures intelligibles, normalisées et statistiques se prêtent bien à la détermination de ce seuil.

$P_2(m) = 0 \quad si \quad la\ détermination\ du\ seuil\ est\ problématique,$
$P_2(m) = 1 \quad si \quad la\ détermination\ du\ seuil\ est\ immédiate.$

2.3.3 Propriété 3 : Mesure non symétrique

Cette propriété est reprise des propriétés T_{a1} (*la mesure m est symétrique*, [TKS02]) et L_3 (*Mesure non symétrique*, [LT04]). En effet, il est préférable qu'une mesure évalue différemment les règles $X \rightarrow Y$ et $Y \rightarrow X$ puisque la prémisse et la conclusion ont des rôles distincts. Néanmoins, dans certains cas, l'orientation du lien entre X et Y peut ne pas fournir des informations supplémentaires à l'utilisateur, i.e., que les mesures appliquées soient symétriques ou non, cela ne changera rien dans les résultats obtenus.

$P_3(m) = 0 \quad si\ \ m\ est\ symétrique \qquad i.e \quad si\ \forall\ X \rightarrow Y \quad m(X \rightarrow Y) = m(Y \rightarrow X)$
$P_3(m) = 1 \quad si\ \ m\ est\ non\ symétrique \quad i.e \quad si\ \exists\ X \rightarrow Y\ /\ m(X \rightarrow Y) \neq m(Y \rightarrow X)$

À titre d'exemple, nous trouvons que la mesure support est une mesure symétrique puisque $support(X \rightarrow Y) = P(XY) = support(Y \rightarrow X)$. Contrairement à la mesure confiance, qui est non symétrique parce que $confiance(X \rightarrow Y) = \frac{P(XY)}{P(X)}$ est différente de $confiance(Y \rightarrow X) = \frac{P(XY)}{P(Y)}$.

2.3.4 Propriété 4 : Mesure non symétrique au sens de la négation de la conclusion

Cette propriété est reprise de la propriété L_2 ($P_4(m) = 1$). Tan et al. [TKS04] expriment une idée similaire avec la deuxième partie de la propriété T_{a3}, la partie indiquant que quelque soit les règles $X \rightarrow Y$, nous devons avoir $m(X \rightarrow Y) = -m(X \rightarrow \overline{Y})$. C'est un cas particulier de $P_4(m) = 1$.

Cette propriété considère ainsi la symétrie au sens de la négation de la conclusion où les exemples d'une règle $X \rightarrow Y$ forment les contre-exemples de l'autre $X \rightarrow \overline{Y}$ [LT04], [Fre99]. Dans le cas où un utilisateur se trouve dans l'obligation d'utiliser une mesure qui n'est pas capable de distinguer entre les exemples de $X \rightarrow Y$ et ceux de $X \rightarrow \overline{Y}$, alors il doit prendre ce comportement en considération lors de l'interprétation de ses résultats.

$P_4(m) = 0 \quad si \quad m\ est\ nc-symétrique \qquad i.e\ \ si\ \ \forall\ X \rightarrow Y \quad m(X \rightarrow Y) = m(X \rightarrow \overline{Y})$
$P_4(m) = 1 \quad si \quad m\ est\ non\ nc-symétrique \quad i.e\ \ si\ \ \exists\ X \rightarrow Y\ /\ m(X \rightarrow Y) \neq m(X \rightarrow \overline{Y})$

Si nous prenons par exemple les deux mesures support et confiance, nous remarquons qu'elles vérifient cette propriété puisque $support(X \rightarrow Y) = P(XY)$ est différent de $support(X \rightarrow \overline{Y}) = P(X\overline{Y})$ et $confiance(X \rightarrow Y) = P(Y/X)$ est différente de $confiance(X \rightarrow \overline{Y}) = P(\overline{Y}/X)$.

2.3.5 Propriété 5 : Mesure évaluant de la même façon $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ dans le cas de l'implication logique

Cette propriété est reprise de la propriété L_4 ($P_5(m) = 1$), où il est préférable d'évaluer semblablement les règles $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ [Kod99], [DRT07].

$$P_5(m) = 0 \quad si \quad \exists\, X \rightarrow Y \;/\; P(Y/X) = 1 \;\; et \;\; m(X \rightarrow Y) \neq m(\overline{Y} \rightarrow \overline{X})$$
$$P_5(m) = 1 \quad si \quad \forall\, X \rightarrow Y \quad P(Y/X) = 1 \;\Rightarrow\; m(X \rightarrow Y) = m(\overline{Y} \rightarrow \overline{X})$$

Par exemple, la mesure support ne vérifie pas cette propriété puisque $support(X \rightarrow Y) \neq support(\overline{Y} \rightarrow \overline{X})$.Par contre, la mesure confiance la vérifie étant donné que $confiance(X \rightarrow Y) = confiance(\overline{Y} \rightarrow \overline{X})$.

2.3.6 Propriété 6 : Mesure croissante en fonction du nombre d'exemples

Cette propriété est reprise des propriétés S_2 (*la mesure m doit être croissante en fonction de* n_{XY}, [PS91a]) et L_6 (*mesure décroissante en fonction du nombre de contre-exemples*, [LT04]), ($P_6(m) = 1$). Elle montre qu'une règle qui a moins de contre-exemples est jugée plus significative. Selon [Fre99], l'intérêt d'une règle peut être mesuré en fonction du nombre élevé de ses exemples ou en fonction du nombre faible de ses contre-exemples. En regardant les définitions des mesures décrites dans l'*annexe A*, nous remarquons que la sémantique de la plupart de ces mesures est dépourvue du nombre de contre-exemples $n_{X\overline{Y}}$. Néanmoins, à cause de la relation suivante entre n_{XY} et $n_{X\overline{Y}}$: $n_{X\overline{Y}} = n_X - n_{XY}$, nous affirmons que ces mesures considèrent implicitement le nombre de contre-exemples.

La *figure 2.2* illustre le comportement d'une mesure m à l'apparition des exemples à proportions (n_X, n_Y et n) fixées.

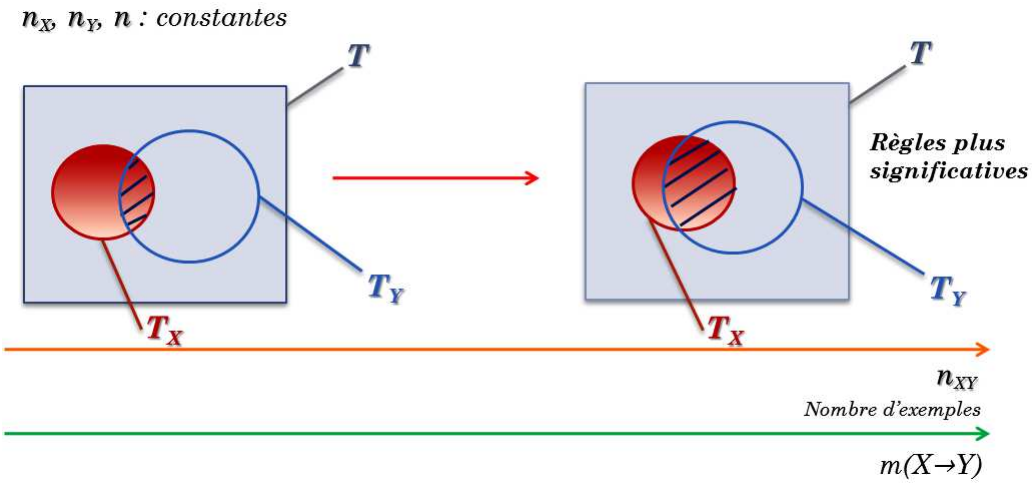

FIGURE *2.2: Influence de l'apparition de* n_{XY} *sur le comportement de* m.

$P_6(m) = 0$ *si* m *n'est pas croissante en fonction de* n_{XY} *i.e. si* $\exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2/$
$n_{X_1} = n_{X_2}$ *et* $n_{Y_1} = n_{Y_2}$ *et* $(n_{X_1}n_{Y_1} < n_{X_2}n_{Y_2}$ *ou* $n_{X_1}n_{\overline{Y}_1} > n_{X_2}n_{\overline{Y}_2})$
et $m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)$

$P_6(m) = 1$ *si* m *est croissante en fonction de* n_{XY} *i.e. si* $\forall X_1 \rightarrow Y_1, \forall X_2 \rightarrow Y_2$
$[n_{X_1} = n_{X_2}$ *et* $n_{Y_1} = n_{Y_2}$ *et* $(n_{X_1Y_1} < n_{X_2Y_2}$ *ou* $n_{X_1\overline{Y}_1} > n_{X_2\overline{Y}_2})]$
$\Rightarrow m(X_1 \rightarrow Y_1) \leq m(X_2 \rightarrow Y_2)$ et
$[\exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2 / n_{X_1} = n_{X_2}$ *et* $n_{Y_1} = n_{Y_2}$ *et* $(n_{X_1Y_1} < n_{X_2Y_2}$ *ou* $n_{X_1\overline{Y}_1} > n_{X_2\overline{Y}_2})$
et $m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)]$

Les deux mesures support et confiance vérifient cette propriété, toutes les deux croissent en fonction du nombre d'exemples n_{XY}.

2.3.7 Propriété 7 : Mesure croissante en fonction de la taille de l'ensemble d'apprentissage

Cette propriété est reprise de la propriété L_{e1} (*la mesure est croissante en fonction de la taille de l'ensemble d'apprentissage*, [Lal02], [GH07]). Elle est également reprise en partie de la propriété T_{a5} (*la mesure* m *doit être invariante lorsque la taille* n *de l'ensemble d'apprentissage* T *augmente et que tous les autres effectifs* n_X, n_Y *et* n_{XY}, *restent constants*), ($P_7 = 0$) puisque Tan et al. [TKS04] préconisent une invariance lorsque la taille de l'ensemble d'apprentissage (*de transactions ne contenant ni la prémisse ni le conséquent*) augmente i.e., lorsque $\forall X_1 \rightarrow Y_1(T_1), \forall X_2 \rightarrow Y_2(T_2)$,

[$n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $n_{X_1Y_1} = n_{X_2Y_2}$ et $n_1 < n_2$] $\Rightarrow m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)$.

C'est donc un cas particulier de $P_7(m) = 0$.

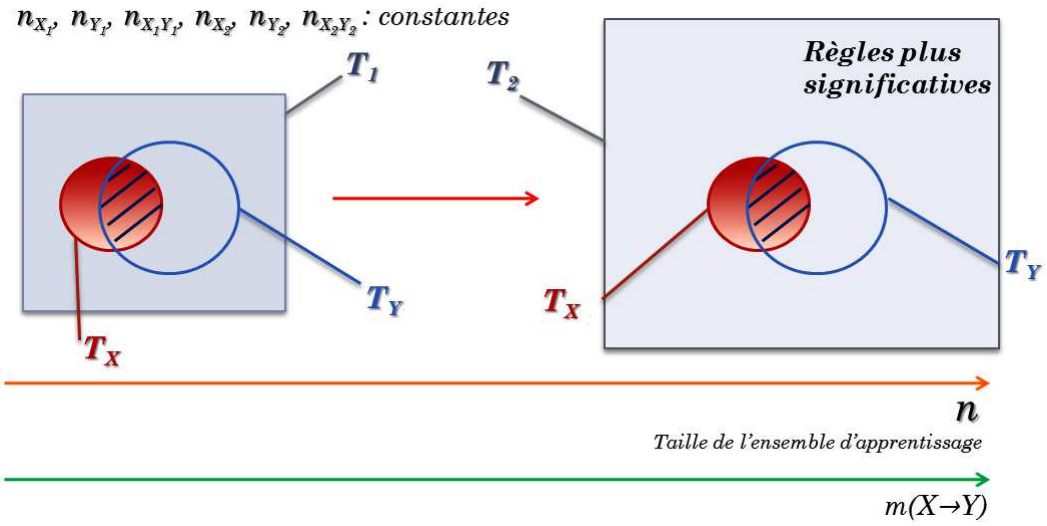

FIGURE *2.3: Illustration de l'influence de n sur le comportement de m.*

La *figure 2.3* illustre comment l'augmentation artificielle de la taille de l'ensemble d'apprentissage intervient statistiquement sur l'évaluation de l'intérêt d'une règle d'association.

Soient T_1 et T_2 deux ensembles d'apprentissage. Soient n_1 la taille du premier ensemble d'apprentissage T_1 et n_2 la taille du second ensemble d'apprentissage T_2. On entend par la notation $X_1 \rightarrow Y_1$ (T_1), la règle $X_1 \rightarrow Y_1$ extraite dans l'ensemble d'apprentissage T_1. Cette règle est d'autant plus significative lorsque T_1 est volumineux. Dès lors, X_1 et Y_1 deviennent rares.

$$
\begin{array}{l}
P_7(m) = 0 \quad si \quad m\ n'est\ pas\ croissante\ en\ fonction\ de\ n \quad i.e. \quad si \\
\exists\, T_1, \exists\, T_2, \exists\, X_1 \rightarrow Y_1\ (T_1), \exists\, X_2 \rightarrow Y_2\ (T_2)\ /\ n_{X_1} = n_{X_2} \quad et \quad n_{Y_1} = n_{Y_2} \\
et \quad n_{X_1Y_1} = n_{X_2Y_2}\ et\ n_1 < n_2 \quad et \quad m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)
\end{array}
$$

$$
\begin{array}{l}
P_7(m) = 1 \quad si \quad m\ est\ croissante\ en\ fonction\ de\ n \quad i.e. \quad si \\
\forall\, T_1, \forall\, T_2, \ \forall\, X_1 \rightarrow Y_1\ (T_1), \forall\, X_2 \rightarrow Y_2\ (T_2), \\
(n_{X_1} = n_{X_2} \quad et \quad n_{Y_1} = n_{Y_2}\ et\ n_{X_1Y_1} = n_{X_2Y_2}\ et\ n_1 < n_2) \\
\Rightarrow \quad m(X_1 \rightarrow Y_1) \leq m(X_2 \rightarrow Y_2) \\
et \quad \exists\, T_1, \exists\, T_2, \exists\, X_1 \rightarrow Y_1\ (T_1), \exists\, X_2 \rightarrow Y_2\ (T_2)\ / \\
(n_{X_1} = n_{X_2} \quad et \quad n_{Y_1} = n_{Y_2}\ et\ n_{X_1Y_1} = n_{X_2Y_2} \quad et \quad n_1 < n_2\ ) \\
et \quad m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)
\end{array}
$$

Comme l'expression des mesures support et confiance est dépourvue de la quantité n, ces deux mesures ne vérifient pas alors cette propriété ($P_7 = 0$) et sont décroissantes en fonction de la taille de l'ensemble d'apprentissage.

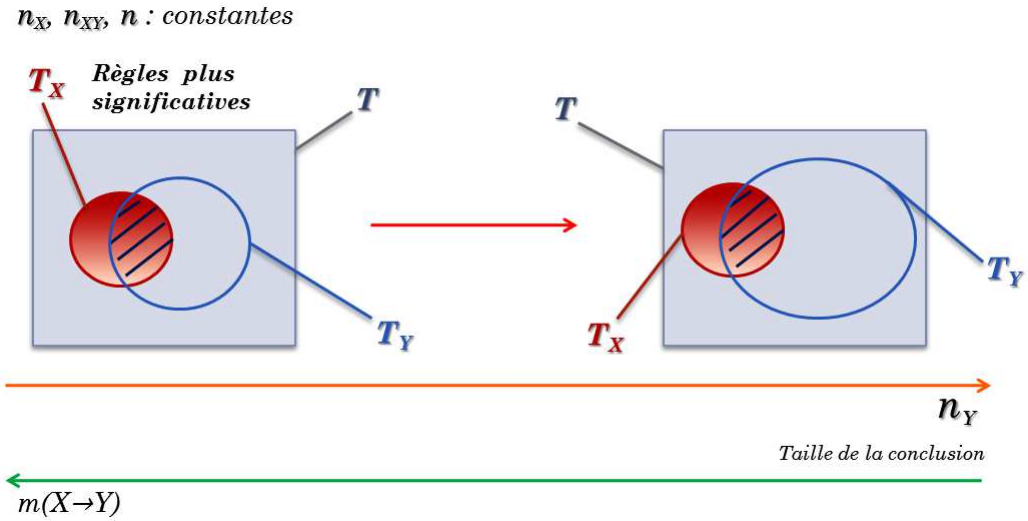

FIGURE *2.4: Illustration de l'influence de n_Y sur le comportement de m.*

2.3.8 Propriété 8 : Mesure décroissante en fonction de la taille du conséquent ou de la taille de la prémisse

Cette propriété est reprise de la propriété L_{10} (*mesure croissante en fonction de la rareté du conséquent*, [LT04]), ($P_8(m) = 1$) et de la propriété S_3 (*la mesure m doit être décroissante en fonction de la taille n_X de la prémisse lorsque le nombre n_{XY} d'exemples et la taille n_Y de la conclusion restent constantes*, [PS91a]).

Étant donnés les nombres d'exemples, de contre-exemples et de la prémisse fixés, nous évaluons l'intérêt d'une règle en mesurant le nombre faible de la taille de sa conclusion. Dans la *figure 2.4*, nous illustrons l'intérêt de la règle par rapport à la taille de la conclusion. Les effectifs n_{XY}, n_X et $n_{X\overline{Y}}$ de la *table de contingence 1.2, page 13* étant fixes, nous remarquons que plus la taille de la conclusion n_Y croît, plus la partie de Y non couverte par X, $n_{\overline{X}Y}$, augmente également. Ainsi, la conclusion est d'autant mieux discernée par la règle que sa taille est petite [PS91a].

$P_8(m) = 0 \quad si \quad m\ n'est\ pas\ décroissante\ en\ fonction\ de\ n_Y \ \ i.e. \quad si \quad \exists\, X_1 \rightarrow Y_1, \exists\, X_2 \rightarrow Y_2\ /$
$n_{X_1} = n_{X_2} \quad et \quad n_{X_1Y_1} = n_{X_2Y_2} \quad et \quad n_{Y_1} < n_{Y_2} \quad et \quad m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)$

$P_8(m) = 1 \quad si \quad m\ est\ décroissante\ en\ fonction\ de\ n_Y \quad i.e. \quad si$
$\forall\, X_1 \rightarrow Y_1, \forall\, X_2 \rightarrow Y_2 \quad (n_{X_1} = n_{X_2} \quad et \quad n_{X_1Y_1} = n_{X_2Y_2} \quad et \quad n_{Y_1} < n_{Y_2})$
$\Rightarrow \quad m(X_1 \rightarrow Y_1) \geq m(X_2 \rightarrow Y_2) \quad et$
$\exists\, X_1 \rightarrow Y_1, \exists\, X_2 \rightarrow Y_2\ /\ n_{X_1} = n_{X_2} \quad et \quad n_{X_1Y_1} = n_{X_2Y_2} \quad et \quad n_{Y_1} < n_{Y_2}$
$et \quad m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)$

Si nous considérons la taille de la prémisse, la propriété $P_8(m) = 1$ s'écrit aussi :

$P_8(m) = 1 \quad si \quad m\ est\ décroissante\ en\ fonction\ de\ n_X \quad i.e. \quad lorsque$
$\forall X_1 \rightarrow Y_1, \forall X_2 \rightarrow Y_2\ (n_{Y_1} = n_{Y_2} \quad et \quad n_{X_1Y_1} = n_{X_2Y_2} \quad et \quad n_{X_1} < n_{X_2})$
$\Rightarrow \quad m(X_1 \rightarrow Y_1) \geq m(X_2 \rightarrow Y_2)$

Le support et la confiance ne vérifient pas cette propriété puisqu'elles ne considèrent pas les règles à faible conclusion.

2.3.9 Propriété 9 : Valeur fixe dans le cas de l'indépendance

Cette propriété est reprise des propriétés L_7 (*valeurs fixes pour l'indépendance*, [LT04]) et S_1 (*la valeur de la mesure m doit être nulle dans le cas de l'indépendance*, [PS91a]). Une mesure m qui vérifie cette propriété $(P_9(m) = 1)$ doit impérativement avoir une valeur fixe a à l'indépendance. Cette caractéristique permettra ainsi de faciliter la tâche de l'utilisateur qui n'a qu'à fixer un seuil de filtre μ qui soit supérieur à la valeur de la mesure pour deux motifs X et Y statistiquement indépendants.

$P_9(m) = 0 \quad si \quad \forall a \in \mathbb{R} \ \exists \ X \rightarrow Y \ / \ P(Y/X) = P(Y) \quad et \quad m(X \rightarrow Y) \neq a$
$P_9(m) = 1 \ (valeur\ fixe) \quad si \quad \exists a \in \mathbb{R} \ \forall X \rightarrow Y \quad P(Y/X) = P(Y) \quad \Rightarrow \quad m(X \rightarrow Y) = a$

La *figure 2.5* présente les valeurs que peut prendre certaines mesures à l'indépendance. Nous reprenons les mesures décrites dans la sous-section 2.2.2, outre les mesures support et confiance.

2.3.10 Propriété 10 : Valeur fixe dans le cas de l'implication logique

Cette propriété est reprise de la propriété L_8 (*valeurs fixes pour l'implication*, [LMP+03], [LT04]). En effet, pour rendre la fixation d'un seuil de filtre facilement gérable par l'utilisateur, il serait plus intéressant qu'une mesure d'intérêt m ait une valeur fixe b à l'implication logique $(P_{10}(m) = 1)$. Dès lors, l'utilisateur n'a qu'à choisir un seuil μ en dessous de la valeur minimale que prend la mesure m en cas d'absence de contre-exemples, pour ne pas écarter ces règles.

Les auteurs dans [LMP+03] montrent que la vérification de cette propriété par une mesure m peut lui causer des problèmes, telle que la perte de son pouvoir discriminant. Cependant, malgré ce défaut, sa vérification reste toujours souhaitable.

$P_{10}(m) = 0 \quad si \quad \forall b \in \mathbb{R} \ \exists \ X \rightarrow Y \ / \ P(Y/X) = 1 \quad et \quad m(X \rightarrow Y) \neq b$
$P_{10}(m) = 1 \ (valeur\ fixe) \quad si \quad \exists b \in \mathbb{R} \ \forall X \rightarrow Y \quad P(Y/X) = 1 \quad \Rightarrow \quad m(X \rightarrow Y) = b$

La *figure 2.5* illustre les valeurs que peut prendre certaines mesures à l'implication logique.

Situation Réf. / Mesure	Équilibre	Indépendance	Implication logique
Confiance	$\frac{1}{2}$	$p(Y)$	1
Confiance causale	$\frac{3}{4} - \frac{P(X)}{4P(\overline{Y})}$	$1 - \frac{1}{2}(P(X) + P(\overline{Y}))$	1
Confiance centrée	$\frac{1}{2} - P(Y)$	0	$1 - P(Y)$
Conviction	$2P(\overline{Y})$	1	$+\infty$
Facteur bayésien	$\frac{P(\overline{Y})}{P(Y)}$	1	$+\infty$
Fiabilité négative	$\frac{\frac{3}{2}P(X) - P(Y)}{P(\overline{Y})}$	$P(\overline{X})$	1
Indice d'implication	$\frac{\sqrt{nP(X)}}{\sqrt{P(\overline{Y})}}(\frac{1}{2} - p(\overline{Y}))$	0	$-\sqrt{nP(X)P(\overline{Y})}$
Intensité d'implication	$P[N(0,1) \geq \sqrt{\frac{nP(X)}{P(\overline{Y})}}(\frac{1}{2} - P(X))]$	$\frac{1}{2}$	1
Intérêt	$\frac{1}{2}P(\overline{Y})$	1	$\frac{1}{P(Y)}$
Loevinger	$\frac{\frac{1}{2} - P(Y)}{P(\overline{Y})}$	0	1
M_{GK}	$\frac{\frac{1}{2} - P(Y)}{1 - P(Y)}$	0	1
Sebag-Schoenauer	1	$\frac{P(Y)}{P(\overline{Y})}$	$+\infty$
Support	$\frac{P(X)}{2}$	$P(X)P(Y)$	$P(X)$
Support sens unique	$\frac{1}{2}log_2(\frac{1}{2P(Y)})$	0	$log_2(P(\overline{Y}))$

FIGURE *2.5: Valeurs des mesures aux trois situations de référence.*

2.3.11 Propriété 11 : Valeur fixe dans le cas de l'équilibre

Cette propriété est reprise de la propriété B_1 (*valeur fixe dans le cas de l'équilibre*, [BGBG5a]). Les auteurs dans [BGBG5a] étudient une autre situation de référence autre que l'indépendance et l'implication logique, qui est l'équilibre. Ils proposent une valeur fixe c à ce point de référence, et ce lorsqu'une règle possède autant d'exemples que de contre-exemples. Ceci permettra à l'utilisateur, comme pour le cas de l'indépendance, de choisir un seuil de filtre μ supérieur à la valeur de la mesure et de ne retenir par conséquent que les règles dépassant cette valeur.

$$P_{11}(m) = 0 \quad si \quad \forall\, c \in \mathbb{R} \;\; \exists\; X \rightarrow Y \;/\; P(Y/X) = P(X)/2 \quad et \quad m(X \rightarrow Y) \neq c$$
$$P_{11}(m) = 1 \;\; (valeur\ fixe) \quad si \quad \exists\, c \in \mathbb{R} \;\; \forall\, X \rightarrow Y \quad P(Y/X) = P(X)/2 \quad \Rightarrow \quad m(X \rightarrow Y) = c$$

De même que pour le cas de l'indépendance et de l'implication logique, la *figure 2.5* illustre les valeurs que peuvent prendre certaines mesures (*définies dans la section 2.2.2*) à l'équilibre.

2.3.12 Propriété 12 : Valeurs identifiables en cas d'attraction entre X et Y

Cette propriété est reprise et généralisée de l'implication I_1 (*les valeurs des mesures doivent être positives en cas d'attraction entre* X *et* Y) de Piatetsky-Shapiro [PS91a], où il serait avan-

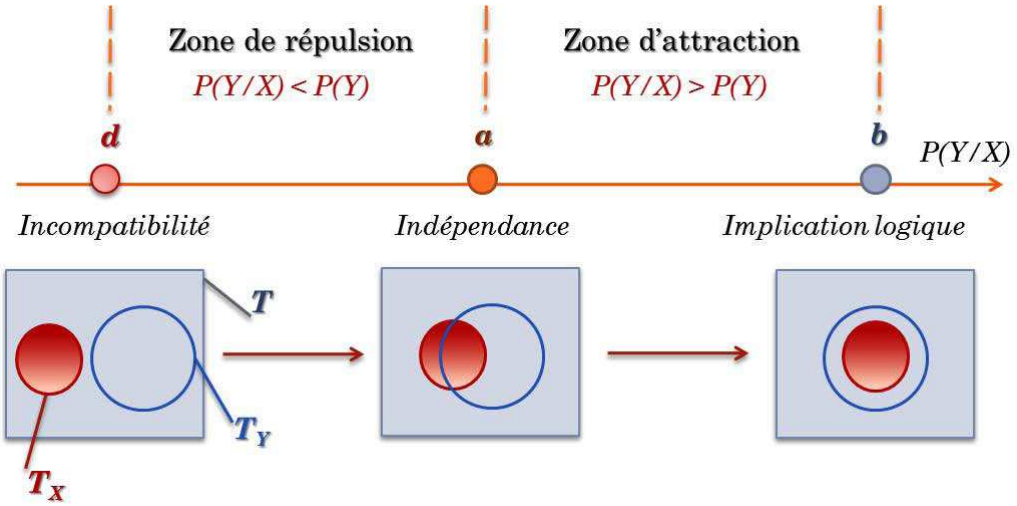

FIGURE *2.6: Identification des zones de répulsion et d'attraction entre deux motifs X et Y.*

tageux pour une mesure d'intérêt qu'elle puisse identifier le niveau d'attraction entre deux motifs X et Y. Dans ce cas, la mesure m doit avoir une valeur fixe à l'indépendance à partir de laquelle deux motifs sont jugées proches. Ainsi, la valeur "a" indiquée dans la *propriété 12* correspond à la valeur fixe dans le cas de l'indépendance. De cette remarque, nous pouvons déduire que si la *propriété 9* n'est pas vérifiée, alors la *propriété 12* ne le sera pas non plus : si $P_9(m) = 0$ alors $P_{12}(m) = 0$. Cette remarque est également valable pour la *propriété 13*.

La zone de droite identifiée dans la *figure 2.6* est la zone d'attraction. Elle présente l'intervalle dans lequel deux motifs X et Y sont positivement corrélés. Cette zone est située entre l'indépendance et l'implication logique.

$$P_{12}(m) = 0 \quad si \quad \forall\, a \in \mathbb{R} \;\; \exists\, X \rightarrow Y \;/\; P(Y/X) > P(Y) \quad et \quad m(X \rightarrow Y) \leq a$$
$$P_{12}(m) = 1 \quad si \quad \exists\, a \in \mathbb{R} \;\; \forall\, X \rightarrow Y \; P(Y/X) > P(Y) \quad \Rightarrow \quad m(X \rightarrow Y) > a$$

Aucune des mesures confiance ou support ne possèdent une valeur fixe à l'indépendance (*figure 2.5*). Par conséquent, elles ne vérifient pas cette propriété ($P_{12}(m) = 0$) puisqu'elles ne peuvent pas identifier le niveau d'attraction entre deux motifs donnés.

2.3.13 Propriété 13 : Valeurs identifiables en cas de répulsion entre X et Y

Cette propriété est reprise et généralisée de l'implication I_2 (*les valeurs des mesures doivent être négatives en cas de répulsion entre X et Y*) de Piatetsky-Shapiro [PS91a], à partir de laquelle nous étudions le niveau de répulsion entre deux motifs. Étant donnée une règle d'association $X \rightarrow Y$, pour qu'une mesure puisse identifier le niveau de répulsion entre X et Y, elle

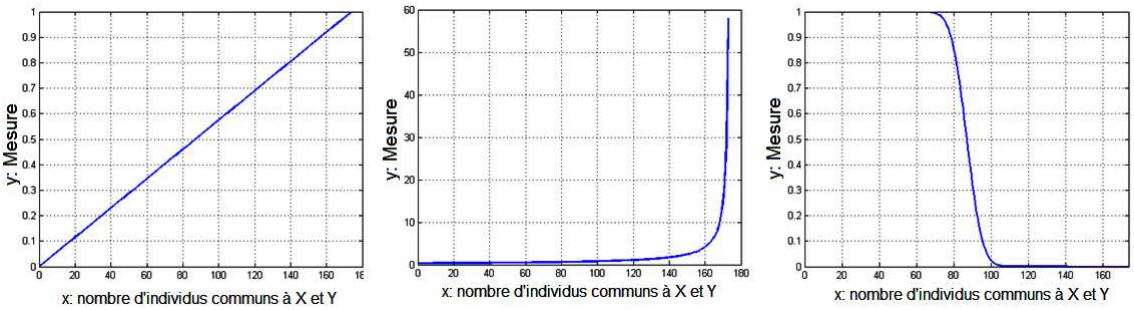

FIGURE *2.7: Différents comportements d'une mesure d'intérêt* m *(linéaire, concave, convexe).*

doit avoir de même que pour la *propriété 12*, une valeur fixe à l'indépendance en dessous de laquelle elle est capable d'identifier à quel point deux motifs sont distants.

La zone de gauche identifiée dans la *figure 2.6* est la zone de répulsion. Elle présente l'intervalle dans lequel deux motifs X et Y sont négativement corrélés. Cette zone est située entre l'incompatibilité et l'indépendance.

$$P_{13}(m) = 0 \quad si \quad \forall\, a \in \mathbf{R} \;\; \exists\, X \rightarrow Y \;/\; P(Y/X) < P(Y) \quad et \quad m(X \rightarrow Y) \geq a$$
$$P_{13}(m) = 1 \quad si \quad \exists\, a \in \mathbf{R} \;\; \forall\, X \rightarrow Y \;\; P(Y/X) < P(Y) \quad \Rightarrow \quad m(X \rightarrow Y) < a$$

De même que pour la propriété 12, les deux mesures confiance et support ne vérifient pas cette propriété ($P_{13}(m) = 0$) puisqu'elles ne possèdent pas une valeur fixe à l'indépendance (*figure 2.5*).

2.3.14 Propriété 14 : Tolérance aux premiers contre-exemples

Cette propriété est reprise de la propriété L_9 (*tolérance aux premiers contre-exemples*, [GKCG01], [LT04]) et nous retenons les trois modalités (*linéarité, concavité, convexité*) proposées par Vaillant dans [Vai06]. À marges n_X et n_Y fixées et en fonction des besoins des utilisateurs, une mesure d'intérêt peut être plus intéressante si elle est capable de tolérer les premiers contre-exemples. Autrement dit, une mesure qui décroît lentement à l'apparition de peu de contre-exemples puis de plus en plus rapidement (*ayant une allure concave*) jusqu'à parvenir à une valeur minimale voir nulle, est préférable [GKCG01]. La *figure 2.7* illustre les trois modalités (*linéarité, concavité, convexité*) de comportement d'une mesure d'intérêt en fonction du nombre d'exemples.

Dans certains domaines d'application tel que le domaine médical, il ne serait pas souhaitable qu'une mesure ait une allure concave mais plutôt convexe afin d'échapper aux périls, contrairement à d'autres applications (*tels que le marketing*), où la prise de risque ne sera pas aussi dommageable pour la société.

$P_{14}(m) = 0 \quad si \quad rejet\ donc\ convexe, \exists\, min_{conf} \in [0,1], \quad \forall\, X_1 \rightarrow Y_1, \forall\, X_2 \rightarrow Y_2, \ \forall\, \lambda \in [0,1]$
$n_{X_1Y_1} \geq min_{conf}\ n_{X_1} \quad et \quad n_{X_2Y_2} \geq min_{conf}\ n_{X_2}$
$impliquent \ \ f_{m,n_{XY}}(\lambda n_{X_1Y_1} + (1-\lambda) n_{X_2Y_2}) \ \leq \ \lambda f_{m,n_{XY}}(n_{X_1Y_1}) + (1-\lambda) f_{m,n_{XY}}(n_{X_2Y_2})$

$P_{14}(m) = 1 \quad si \quad indiffférence\ notamment\ linéaire \quad i.e. \quad P_{14}(m) \neq 0 \quad et \quad P_{14}(m) \neq 2$

$P_{14}(m) = 2 \quad si \quad tolérance\ alors\ concave\ \exists\, min_{conf} \in [0,1], \quad \forall\, X_1 \rightarrow Y_1, \forall\, X_2 \rightarrow Y_2, \ \forall\, \lambda \in [0,1]$
$n_{X_1Y_1} \geq min_{conf}\ n_{X_1} \quad et \quad n_{X_2Y_2} \geq min_{conf}\ n_{X_2}$
$impliquent \ \ f_{m,n_{XY}}(\lambda n_{X_1Y_1} + (1-\lambda) n_{X_2Y_2}) \ \geq \ \lambda f_{m,n_{XY}}(n_{X_1Y_1}) + (1-\lambda) f_{m,n_{XY}}(n_{X_2Y_2})$

La notation $f_{m,n_{XY}}$ correspond à la fonction d'évolution de la mesure m en fonction de n_{XY} lorsque les effectifs n_X, n_Y et n restent constants.

Le support et la confiance sont toutes les deux linéaires en fonction des contre-exemples $n(X\overline{Y})$, indifférentes aux premiers contre-exemples.

2.3.15 Propriété 15 : Invariance en cas de dilatation de certains effectifs

Cette propriété est reprise de la propriété T_{a2} de [TKS02]. À proportions fixées, il serait préférable qu'une mesure soit invariante en cas de dilatation de certains effectifs (n_{XY}, $n_{\overline{XY}}$, $n_{X\overline{Y}}$ et $n_{\overline{X}Y}$).

$P_{15}(m) = 0 \ \ (variance) \quad si \quad \exists\, (k_1,k_2) \in \mathbb{N}^{*2}, \exists\, X_1 \rightarrow Y_1, \exists\, X_2 \rightarrow Y_2$
$[n_{X_1Y_1} = k_1\, n_{X_2Y_2} \quad et \quad n_{X_1\overline{Y}_1} = k_1\, n_{X_2\overline{Y}_2} \quad et \quad n_{\overline{X}_1Y_1} = k_2\, n_{\overline{X}_2Y_2} \quad et \quad n_{\overline{X}_1\overline{Y}_1} = k_2\, n_{\overline{X}_2\overline{Y}_2}$
$et \quad m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)]$
$ou \quad [n_{X_1\overline{Y}_1} = k_1\, n_{X_2\overline{Y}_2} \quad et \quad n_{\overline{X}_1\overline{Y}_1} = k_1\, n_{\overline{X}_2\overline{Y}_2} \quad et \quad n_{X_1Y_1} = k_2\, n_{X_2Y_2} \quad et \quad n_{\overline{X}_1Y_1} = k_2\, n_{\overline{X}_2Y_2}$
$et \quad m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)]$

$P_{15}(m) = 1 \ \ (invariance) \quad si \quad \forall\, (k_1,k_2) \in \mathbb{N}^{*2}, \forall\, X_1 \rightarrow Y_1, \forall\, X_2 \rightarrow Y_2$
$[(n_{X_1Y_1} = k_1 n_{X_2Y_2} \quad et \quad n_{X_1\overline{Y}_1} = k_1 n_{X_2\overline{Y}_2} \quad et \quad n_{\overline{X}_1Y_1} = k_2 n_{\overline{X}_2Y_2} \quad et \quad n_{\overline{X}_1\overline{Y}_1} = k_2 n_{\overline{X}_2\overline{Y}_2})$
$\Rightarrow \ \ m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)]$
$et \ \ [(n_{X_1\overline{Y}_1} = k_1 n_{X_2\overline{Y}_2} \quad et \quad n_{\overline{X}_1\overline{Y}_1} = k_1 n_{\overline{X}_2\overline{Y}_2} \quad et \quad n_{X_1Y_1} = k_2 n_{X_2Y_2} \quad et \quad n_{\overline{X}_1Y_1} = k_2 n_{\overline{X}_2Y_2})$
$\Rightarrow \ \ m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)]$

Il est à noter que la formalisation de cette propriété par [TKS02] à l'aide de matrices est plus compacte que ce que nous vous présentons. Cependant, nous recherchons dans le présent chapitre une formalisation uniforme pour toutes les propriétés.

La confiance et le support sont deux mesures qui varient en fonction de la croissance de certains effectifs. Par conséquent, elles ne vérifient pas cette propriété ($P_{15}(m) = 0$).

2.3.16 Propriété 16 : Relation souhaitée entre les règles $X \rightarrow Y$ et $\overline{X} \rightarrow Y$

Cette propriété est reprise de la première partie de la propriété T_{a3}, où une mesure m qui est capable de différencier entre les règles $X \rightarrow Y$ et $\overline{X} \rightarrow Y$ selon une relation d'opposition serait souhaitable.

$$P_{16}(m) = 0 \quad si \quad \exists\, X \rightarrow Y \quad m(X \rightarrow Y) \neq -m(\overline{X} \rightarrow Y)$$
$$P_{16}(m) = 1 \quad si \quad \forall\, X \rightarrow Y \quad m(X \rightarrow Y) = -m(\overline{X} \rightarrow Y)$$

Cette propriété n'est pas vérifiée par les mesures confiance et support puisque $confiance(\overline{X} \rightarrow Y) \neq -confiance(X \rightarrow Y)$ et $support(\overline{X} \rightarrow Y) \neq -support(X \rightarrow Y)$.

2.3.17 Propriété 17 : Relation souhaitée entre les règles $X \rightarrow Y$ et $X \rightarrow \overline{Y}$

Cette propriété est reprise de la deuxième partie de la propriété T_{a3} où une mesure m qui différencie entre les règles $X \rightarrow Y$ et $X \rightarrow \overline{Y}$ selon une relation d'opposition serait préférable.

$$P_{17}(m) = 0 \quad si \quad \exists\, X \rightarrow Y \quad m(X \rightarrow Y) \neq -m(X \rightarrow \overline{Y})$$
$$P_{17}(m) = 1 \quad si \quad \forall\, X \rightarrow Y \quad m(X \rightarrow Y) = -m(X \rightarrow \overline{Y})$$

Nous avons $confiance(X \rightarrow \overline{Y}) \neq -confiance(X \rightarrow Y)$ et $support(X \rightarrow \overline{Y}) \neq -support(X \rightarrow Y)$. Par conséquent, ces deux mesures ne permettent pas de différencier entre les règles $X \rightarrow Y$ et $X \rightarrow \overline{Y}$, d'où $P_{17}(m) = 0$.

2.3.18 Propriété 18 : Relation souhaitée entre les règles $X \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$

Cette propriété est reprise de la propriété T_{a4} de [TKS02]. Les auteurs expliquent qu'il serait avantageux pour une mesure m d'avoir une même valeur pour ces deux types de règles $X \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$.

$$P_{18}(m) = 0 \quad si \quad \exists\, X \rightarrow Y \;/\; m(X \rightarrow Y) \neq m(\overline{X} \rightarrow \overline{Y})$$
$$P_{18}(m) = 1 \quad si \quad \forall\, X \rightarrow Y \quad m(X \rightarrow Y) = m(\overline{X} \rightarrow \overline{Y})$$

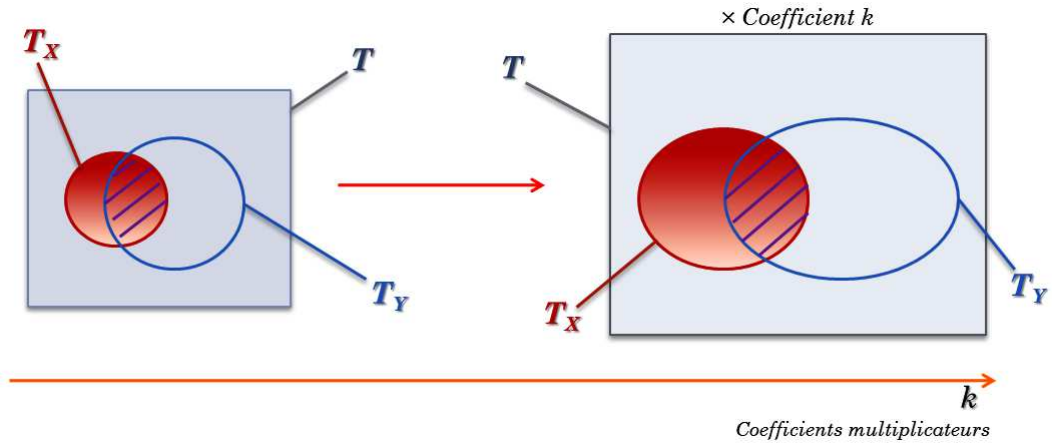

FIGURE *2.8: Illustration du comportement statistique d'une mesure* m.

De même que pour les deux propriétés précédentes, les deux mesures confiance et support ne valident pas cette propriété ($P_{18}(m) = 0$).

2.3.19 Propriété 19 : Taille de la prémisse fixe ou aléatoire

Cette propriété est issue de la propriété L_5 (*taille de la prémisse fixe ou aléatoire*, [LT04]). Une mesure m qui est fondée sur l'un de ces modèles probabilistes : distribution normale, binomiale, poisson ou encore hypergéométrique, possède nécessairement une taille aléatoire de la prémisse.

$$P_{19}(m) = 0 \quad (taille\ fixe) \qquad si \quad m\ n'est\ pas\ fondée\ sur\ un\ modèle\ probabiliste$$
$$P_{19}(m) = 1 \quad (taille\ aléatoire) \quad si \quad m\ est\ fondée\ sur\ un\ modèle\ probabiliste$$

Le support et la confiance ne sont fondés sur aucun modèle probabiliste. Ainsi, la taille de la prémisse pour ces deux mesures est fixe.

2.3.20 Propriété 20 : Mesure descriptive ou statistique

Cette propriété est issue de la propriété L_{11} (*mesure descriptive ou statistique*, [LT04]). Étant donnés tous les effectifs de la *table de contingence 1.2, page 13* fixes, une mesure est dite *descriptive* si elle est invariante à la croissance du nombre d'enregistrements. Contrairement aux mesures dites *statistiques*, qui sont sensibles (*croissantes*) à l'augmentation des données. Ces dernières sont plus appréciables puisqu'une règle est d'autant plus significative que le nombre d'enregistrements n est grand [Gra79], [Lal02], [GCB+04], [Vai06]. La *figure 2.8* illustre ce phénomène.

$P_{20}(m) = 0$ $(descriptive\ ou\ invariante)$ si $\forall\, k \in \mathbb{N}^*, \forall\, X_1 \rightarrow Y_1, \forall\, X_2 \rightarrow Y_2,$
$(n_{X_1Y_1} = k\, n_{X_2Y_2}$ et $n_{X_1\overline{Y}_1} = k\, n_{X_2\overline{Y}_2}$ et $n_{\overline{X}_1Y_1} = k\, n_{\overline{X}_2Y_2}$ et $n_{\overline{X}_1\overline{Y}_1} = k\, n_{\overline{X}_2\overline{Y}_2})$
$\Rightarrow$ $m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)$

$P_{20}(m) = 1$ $(statistique\ ou\ variante)$ si $\exists\, k \in \mathbb{N}^*, \exists\, X_1 \rightarrow Y_1, \exists\, X_2 \rightarrow Y_2$
$(n_{X_1Y_1} = k\, n_{X_2Y_2}$ et $n_{X_1\overline{Y}_1} = k\, n_{X_2\overline{Y}_2}$ et $n_{\overline{X}_1Y_1} = k\, n_{\overline{X}_2Y_2}$ et $n_{\overline{X}_1\overline{Y}_1} = k\, n_{\overline{X}_2\overline{Y}_2})$
et $m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)$

Cependant, nous signalons que pour certaines mesures (*statistiques*), essentiellement celles qui sont bornées, elles risquent d'atteindre leur valeur maximale et par conséquent risquent de perdre leur pouvoir discriminant en considérant une taille assez élevée des données. Ces dernières doivent ainsi évaluer un nombre important de règles, des évaluations qui peuvent être proches du maximum. Inversement à d'autres mesures (*descriptives*), qui sont indépendantes de la taille des données.

Selon Lallich et Teytaud [LT04], la taille de l'ensemble d'apprentissage n doit intervenir dans l'évaluation d'une mesure m.

Par exemple, le support et la confiance sont deux mesures descriptives puisqu'elles sont indépendantes de la taille des données. Par conséquent, elles ne vérifient pas cette propriété.

2.3.21 Propriété 21 : Mesure discriminante

Cette propriété est issue de la propriété L_{12} (*mesure discriminante*, [LT04]). Une mesure est dite discriminante si elle est capable de distinguer les règles intéressantes lorsque la taille de l'ensemble d'apprentissage n croit. Autrement dit, une mesure permet de restituer des valeurs distinctes aux règles pour des niveaux d'implication différents [GKCG01].

$P_{21}(m) = 0$ $(non\ discriminante)$ si $\exists\, \eta \in \mathbb{N}^*$ $\forall\, n > \eta$ $\forall\, X_1 \rightarrow Y_1, \forall\, X_2 \rightarrow Y_2$
$[P(Y_1/X_1) > P(Y_1)$ et $P(Y_2/X_2) > P(Y_2)]$ $\Rightarrow$ $m(X_1 \rightarrow Y_1) \simeq m(X_2 \rightarrow Y_2)$
$P_{21}(m) = 1$ $(discriminante)$ si $\forall\, \eta \in \mathbb{N}^*$ $\exists\, n > \eta$ $\exists\, X_1 \rightarrow Y_1$ $\exists\, X_2 \rightarrow Y_2$
$P(Y_1/X_1) > P(Y_1)$ et $P(Y_2/X_2) > P(Y_2)$ et $m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)$

Le support et la confiance sont deux mesures discriminantes capables de différencier entre les règles.

2.3.22 Propriété 22 : Mesure robuste

Cette propriété est issue de la propriété Br_1 (*mesure robuste*, [BLL12]). Idéalement, une mesure d'intérêt doit prendre en considération le bruit existant dans les données [AK02]. Ce bruit peut être identifié sous différentes formes, e.g., il pourrait s'agir d'une faute de frappe pendant la création de la base ou une valeur qui est absente dans les données, ou l'introduction de nouvelles transactions, etc. Si le bruit est identifié par la mesure d'intérêt, alors la mesure est dite robuste ($P_{22} = 1$).

$$P_{22}(m) = 0 \quad si \quad m\ n'est\ pas\ robuste,$$
$$P_{22}(m) = 1 \quad si \quad m\ est\ robuste$$

Comme il est très difficile, voire même rare, d'avoir des données réelles "*parfaites*", le traitement des données bruitées s'avère alors indispensable afin de donner des informations sécurisées à l'utilisateur. Certes, l'étude de cette propriété pour un nombre important de mesures nécessite des outils de calcul très poussés afin d'éviter les erreurs de calcul [BMLL10a], la raison pour laquelle nous ne la considérons pas lors de l'étude des mesures.

Selon Le Bras [Bra11], les mesures support et confiance sont jugées deux mesures robustes.

Après avoir formalisé les propriétés, nous allons les étudier sur les différentes mesures existantes dans la littérature (*tables 2.1 et 2.2*).

2.4 Évaluation des mesures d'intérêt selon les propriétés

Dans le but d'analyser le comportement des différentes mesures d'intérêt objectives rappelées en *annexe A*, nous étudions dans ce qui suit, la présence ou l'absence des propriétés mises en évidence dans le *chapitre 1* et formalisées dans la *section 2.3*. Ce travail va déboucher sur la construction d'une matrice, que nous illustrons dans les *tables 2.3 et 2.4*. Cette matrice étudie les *61* mesures d'intérêt définies dans l'*annexe A*. Nous notons par "I", la valeur indéfinie de la propriété P_{14} sur la mesure *(16) Dépendance pondérée*. Cette valeur dépend des paramètres k et m.

Toutefois, parmi les 22 propriétés mises en évidence dans la *section 2.3*, seulement 19 sont étudiées. En effet, les propriétés "P_1 *: intelligibilité ou compréhensibilité de la mesure*", "P_2 *: facilité à fixer un seuil d'acceptation de la règle*" et "P_{22} *: Mesure robuste*" n'ont pas été retenues dans cette étude car nous pensons que les deux premières sont subjectives et

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Mesures
0	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	(1) Coefficient corrélation
0	1	1	1	1	1	1	0	0	1	1	1	0	0	0	1	0	0	1	(2) Cohen
1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	(3) Confiance
1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	(4) Confiance causale
1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	1	(5) Pavillon
1	1	1	1	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1	(6) Ganascia
1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	(7) Confiance confirmée causale
1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	(8) Confirmation causale
1	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	(9) Confirmation descriptive
1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	(10) Conviction
0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	(11) Cosinus
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	(12) Couverture
0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	(13) Czekanowski-dice
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	(14) Dépendance
1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	(15) Dépendance causale
1	1	0	0	1	0	0	0	0	0	0	I	0	0	0	0	0	0	1	(16) Dépendance pondérée
1	1	0	1	1	1	1	0	0	1	1	0	1	0	0	0	0	0	1	(17) Facteur bayésien
1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	1	(18) Loevinger
1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	(19) Fiabilité négative
0	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	1	(20) Force collective
1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	(21) Fukuda
0	1	0	1	1	1	1	0	0	1	1	2	0	0	0	0	0	0	1	(22) Gain informationnel
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	(23) Gini
0	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	0	0	1	(24) Goodman
1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	(25) Indice d'implication
1	1	1	1	0	0	0	0	1	0	0	2	0	0	0	0	1	1	0	(26) IPEE
1	1	1	1	0	0	0	0	1	0	0	2	0	0	0	0	1	1	1	(27) IP3E
1	1	1	1	0	1	0	0	0	0	0	2	1	0	0	0	1	1	1	(28) IPD
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	(29) Information mutuelle
1	1	1	1	1	1	1	0	0	1	1	2	1	0	0	0	1	1	0	(30) II
1	1	1	1	1	0	0	0	0	0	0	2	0	0	0	0	1	1	1	(31) IIE
1	1	1	1	1	0	1	0	0	1	1	2	0	0	0	0	1	1	1	(32) IIER
0	1	1	1	1	1	1	0	0	1	1	2	0	0	0	0	1	1	0	(33) IVL
0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1	(34) Intérêt
0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	(35) Jaccard
1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	(36) J-Mesure
1	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1	(37) Klosgen

TABLE *2.3: Matrice décrivant les mesures d'intérêt selon les propriétés(1).*

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Mesures
0	**1**	0	**1**	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	**1**	(38) Kulczynski
1	**1**	**1**	**1**	0	0	0	0	**1**	0	0	**1**	0	0	0	0	0	0	**1**	(39) Laplace
1	**1**	0	**1**	**1**	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	(40) Leverage
1	**1**	**1**	**1**	**1**	0	**1**	**1**	0	**1**	**1**	**1**	0	0	**1**	0	0	0	**1**	(41) M_{GK}
1	**1**	0	**1**	0	0	0	0	**1**	0	0	**1**	0	0	0	0	0	0	**1**	(42) Moindre contradiction
0	**1**	**1**	**1**	**1**	**1**	**1**	0	0	**1**	**1**	**1**	0	**1**	**1**	**1**	0	0	**1**	(43) Nouveauté
0	0	**1**	0	0	0	**1**	0	0	**1**	0	**1**	0	0	0	**1**	0	0	**1**	(44) Pearl
0	**1**	**1**	**1**	**1**	**1**	**1**	0	0	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	(45) Piatetsky-Shapiro
0	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	**1**	0	0	0	**1**	0	0	**1**	(46) Précision
1	**1**	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	(47) Prévalence
0	**1**	**1**	**1**	**1**	0	**1**	**1**	0	**1**	**1**	**2**	**1**	**1**	**1**	**1**	0	0	**1**	(48) Q de Yule
1	**1**	0	**1**	0	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	(49) Rappel
0	**1**	**1**	**1**	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0	0	**1**	0	0	**1**	(50) Ratio des chances
1	**1**	0	**1**	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	0	0	**1**	(51) Risque relatif
1	**1**	0	**1**	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	**1**	(52) Sebag
1	**1**	0	**1**	**1**	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	(53) Spécificité
0	**1**	0	**1**	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	**1**	(54) Support
1	**1**	0	0	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	0	0	**1**	(55) Support sens unique
0	**1**	0	0	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0	0	0	0	0	**1**	(56) Support double sens
1	**1**	**1**	**1**	0	0	0	**1**	**1**	0	0	**2**	0	0	0	0	0	0	0	(57) Taux d'exemples
0	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	**1**	0	**1**	**1**	**1**	**1**	0	**1**	(58) VT100
0	0	**1**	0	0	0	**1**	0	0	**1**	0	0	0	0	0	**1**	0	0	**1**	(59) Variation support double sens
0	**1**	**1**	**1**	**1**	0	**1**	**1**	0	**1**	**1**	0	**1**	**1**	**1**	**1**	0	0	**1**	(60) Y de Yule
1	**1**	**1**	**1**	**1**	0	**1**	**1**	0	**1**	**1**	**2**	**1**	0	**1**	0	0	0	**1**	(61) Zhang

TABLE *2.4: Matrice décrivant les mesures d'intérêt selon les propriétés(2).*

fonction de la connaissance qu'a l'utilisateur en statistique. Vaillant [Vai06] a également souligné ce point dans sa thèse. Nous pensons aussi que la troisième propriété est difficile à étudier puisqu'elle nécessite des outils de calcul très poussés afin d'éviter les erreurs de calcul, comme le mentionne Le Bras dans [Bra11].

Pour remplir la matrice d'évaluation, nous procédons de la manière suivante : si une mesure vérifie une propriété alors nous mettons 1 dans sa case correspondante, sinon nous mettons 0. Seule la *propriété 14* possède 3 valeurs selon l'allure de l'évolution de la courbe en fonction du nombre d'exemples, nous mettons ou bien 0 (*convexe*), ou bien 1 (*linéaire*) ou bien 2 (*concave*).

Dans ce qui suit, nous exposons les relations mathématiques qui peuvent exister entre les différentes mesures d'intérêt.

2.5 Relations mathématiques entre les mesures

Il serait intéressant d'observer les relations mathématiques qui peuvent exister entre les mesures d'intérêt. Ceci pourrait être utile lors de l'étude du comportement des mesures, essentiellement dans le cas où nous identifions des relations de proportionnalité. La *table 2.5* présente quelques relations découvertes suite à l'examen des définitions des mesures. Ce travail peut en effet être aussi utile pour réduire le nombre de mesures étudiées dans le cas où nous proposons un cadre empirique pour l'analyse des mesures d'intérêt. Par exemple, par

l'identification de forte dépendance entre deux mesures, comme c'est le cas des mesures *Intérêt* et *Gain informationnel* (*ligne 23, table 2.5*), qui sont liées par une relation logarithmique, ou encore des mesures *Piatetsky-shapiro* et *Nouveauté* (*ligne 27, table 2.5*) qui sont proportionnelles, nous pouvons se restreindre à l'étude d'une seule des deux mesures. Si la valeur de *Piatetsky-shapiro* augmente, alors la valeur de *Nouveauté* augmente également, encore, si la valeur de *Conviction* augmente, c'est aussi le cas de la mesure *Facteur de certitude* (*ligne 18, table 2.5*).

2.6 Conclusion

Ce chapitre a résumé les différentes mesures d'intérêt rencontrées dans la littérature. Il propose une formalisation de "*bonnes*" propriétés pour leur description. Cette formalisation est essentielle afin d'éliminer d'éventuelles interprétations, comme par exemple considérer une croissance stricte des propriétés P_6, P_7 et P_8, pouvant engendrer la construction de matrices différentes. S'appuyant sur ces propriétés, une étude formelle est alors réalisée sur 61 mesures objectives afin d'en fournir une caractérisation. Cette étude permet la construction d'une matrice d'évaluation des mesures et porte sur 19 propriétés parmi les 22 recensées dans la littérature. Seules 3 propriétés posent des difficultés d'interprétation.

Ce travail est le point de départ pour une catégorisation des mesures en vue d'aider l'utilisateur dans le choix de ses mesures, dans la phase de post-traitement en fouille de données. Il permettra également la recherche et l'identification des propriétés redondantes.

$N\circ$	$Formule$
1	Pearl $=$ Couverture $\times$ | Confiance centrée |
2	Information Mutuelle $= \frac{VS(XY)}{-P(X)log_2P(X)-P(\overline{X})log_2P(\overline{X})}$
3	Dépendance $=$ | Confiance $-$ Prévalence |
4	Laplace $= \frac{Confiance\times(n\times P(XY)+1)}{n\times P(XY)+2Confiance}$
5	Sebag $= \frac{1}{\frac{1}{Confiance}-1}$
6	Moindre contradiction $=$ Taux d'exemples $\times$ Rappel
7	Ganascia $= 2\times Confiance-1$
8	Confirmation descriptive $=$ Couverture $\times(Ganascia-2)$
9	Cosinus $= \frac{Support}{\sqrt{Couverture\times Prevalence}}$
10	Czekanowski-dice $= \frac{2Support}{Couverture+Prevalence}$
11	Jaccard $= \frac{1}{\frac{Couverture-Support}{Support}+\frac{1}{Rappel}}$
12	Kulczynski $= \frac{Support\times Jaccard}{Support-(Couverture\times Jaccard)}$
13	Spécificité $= Confiance(\overline{X}\rightarrow\overline{Y})$
14	Fiabilité négative $= Confiance(\overline{Y}\rightarrow\overline{X})$
15	Confirmation causale $=$ Précision $-2(Couverture-Support)$
16	Confiance confirmée causale $=$ Confiance causale $+$ Confiance -1
17	Confiance causale $= \frac{1}{2}\left(1+Confiance(X\rightarrow Y)+Confiance(\overline{Y}\rightarrow X)\right)$
18	Facteur de certitude $= 1-\frac{1}{Conviction}$
19	Klosgen $= \sqrt{support}\times$ Confiance centrée
20	Support sens unique $= \frac{Support\ double\ sens}{Couverture}$
21	Risque Relatif $= \frac{P(\overline{Y})Facteur\ de\ certitude+P(Y)}{P(Y/\overline{X})}$
22	Facteur bayésien $=$ Conviction $\times$ Intérêt
23	Gain informationnel $= log_2$(Intérêt)
24	Q de Yule $= \frac{Ratio\ des\ chances-1}{Ratio\ des\ chances+1}$
25	Y de Yule $= \frac{\sqrt{Ratio\ des\ chances}-1}{\sqrt{Ratio\ des\ chances}+1}$
26	$Si\ \ P(Y/X)\geqslant P(Y)\ \ alors\ \ Zhang = \frac{P(X)P(\overline{Y})}{M_{GK}\times P(X\overline{Y})P(Y)\times max(facteurbayesien,1)}$ $sinon\ Zhang = \frac{P(X)}{M_{GK}\times P(X\overline{Y})\times max(Facteur\ bayesien,1)}$
27	Piatestsky-shapiro $= n$| Nouveauté |
28	Taux d'exemples $= 2-\frac{1}{Confiance} = 1-\frac{1}{Sebag}$
29	$Si\ P(Y/X)\geqslant P(Y)\ alors\ M_{GK} = Facteur\ de\ certitude\ sinon\ M_{GK} = \frac{p(\overline{Y})}{p(Y)}\times Facteur\ de\ certitude$
30	Pavillon $= P(\overline{Y})\times Facteur\ de\ certitude$
31	Klosgen $= \sqrt{P(X)}\times Pavillon$

TABLE *2.5: Certaines relations mathématiques entre les mesures d'intérêt*

Points clés

Positionnement :

– *Étude théorique d'une soixantaine de mesures d'intérêt selon un nombre important de propriétés formelles.*

Contribution :

– *Formalisation des propriétés des mesures ;*
– *Évaluation de* 61 *mesures selon* 19 *propriétés.*

Publications :

– *S. Guillaume and D. Grissa and E. Mephu Nguifo (2010). Propriétés des mesures d'intérêt pour l'extraction des règles*. Dans Actes de l'atelier QDC de la conférence EGC, pages 15–28, Hammamet, Tunisie.

CHAPITRE 3

Classification des mesures d'intérêt : méthode sans recouvrement

Sommaire

3.1 Introduction

De nombreuses mesures d'intérêt existent dans la littérature pour faire face aux limites de l'approche support-confiance. Ce nombre important de mesures met l'utilisateur dans une situation difficile quant à la sélection d'une ou plusieurs mesure(s) complémentaire(s) capable(s) d'éliminer les règles non pertinentes extraites par le couple (*support, confiance*). Ainsi, et afin d'aider l'utilisateur dans le choix d'un bon ensemble de mesures d'intérêt qui répond à ses besoins, nous souhaitons détecter des groupes de mesures avec des propriétés similaires. D'où l'objectif principal de ce chapitre, qui est de proposer des classes ou groupes de mesures qui vont permettre à l'utilisateur, d'une part, de restreindre le nombre de mesures à choisir, et d'autre part, d'orienter son choix en fonction des propriétés qu'il souhaiteraient que ces mesures vérifient.

Ce travail s'appuie sur l'étude formelle réalisée dans le chapitre précédent sur les mesures et leurs propriétés, dont résulte une matrice d'évaluation de 61 mesures sur 19 propriétés. Étant donnée cette matrice, nous cherchons à identifier des classes de mesures ayant des comportements similaires par rapport à l'ensemble des propriétés que nous avons dégagées précédemment. Toutefois, nous ne cherchons à expliquer ni les propriétés ni les mesures répertoriées dans la littérature puisqu'elles peuvent être trouvées dans les travaux de synthèse [TKS02], [LT04], [GH07], [Fen07] et [Vai06]. L'identification de ces classes de mesures est effectuée en utilisant des techniques bien connues en classification non supervisée comme la méthode de classification ascendante hiérarchique et la méthode de partitionnement des *k-moyennes*. Un consensus sera dégagé à partir des résultats obtenus avec ces deux techniques. Néanmoins, avant de lancer cette recherche de classes, il nous semble essentiel de vérifier si la matrice de 61 mesures $\times$ 19 propriétés (*identifiée dans le chapitre précédent*) ne peut être simplifiée en supprimant d'éventuelles duplications de lignes ou de colonnes.

Ce chapitre est organisé comme suit. La section 3.2 définit brièvement la classification non supervisée et expose ses objectifs, les méthodes et les critères choisis. La section 3.3 étudie la matrice d'évaluation des mesures selon les propriétés, en vérifiant si celle-ci ne pourrait être simplifiée, et restitue les résultats de la classification obtenue par les deux techniques de classification choisies. La section 3.4 propose une sémantique aux classes extraites. Pour finir, la section 3.5 valide la classification retenue en comparaison avec celles dégagées, respectivement, par Vaillant [Vai06], Huynh et al. [HGB+07], Heravi et Zaiane [HZ10], Le bras [Bra11], Lesot et Rifqi [LR10], et Zighed et al. [ZAB11].

3.2 Classification non supervisée

L'objectif de la classification déborde le cadre strictement exploratoire. Il s'agit de rechercher une typologie, ou segmentation, par la répartition des individus en classes, ou catégories, à partir de traits descriptifs (*attributs, caractéristiques, etc.*). On distingue essentiellement deux types de classification : supervisée et non-supervisée.

Dans le cadre de ce chapitre, nous nous intéressons au problème de la classification non supervisée pour catégoriser les mesures d'intérêt. Cette méthodologie, à l'inverse des méthodes de classification supervisée (*ou catégorisation*), ne présuppose pas une connaissance a priori de la structure du corpus.

3.2.1 Préparation des données

Avant de procéder à la classification, il serait important de préciser la nature des données (*ou des variables*). Ces dernières, à caractère quantitatif ou qualitatif, peuvent prendre différentes formes. Les variables à caractère quantitatif se subdivisent eux-mêmes en deux espèces : continues et discrètes. Les variables à caractère qualitatif sont des données descriptives, qui peuvent être sous forme de variables qualitatives nominales ou ordinales. Une définition détaillée des différents types de données est présentée en *annexe B*.

Les méthodes de classification sont sensibles aux types des données manipulées. Nous retrouvons par exemple que certaines méthodes sont mises en défaut par les variables continues alors que d'autres peuvent être sensibles à la présence de variables discrètes. Ainsi, il faut prendre ce critère en considération et adapter les variables d'entrée aux méthodes choisies. D'où l'étape de transformation des données de l'ECD.

Dans le reste de ce manuscrit, nous considérons les variables qualitatives nominales.

3.2.2 Objectifs de la classification

L'objectif de la classification non supervisée [Har75], [JD88], également appelée clustering en anglais [1], est de découvrir des formes cachées de l'ensemble des individus, ces formes étant des groupes ou classes. Pour ce faire, il s'agit de fractionner l'ensemble hétérogène d'individus à l'intérieur d'une population (*l'ensemble des enregistrements*) en un certain nombre k de sous-ensembles plus homogènes, appelées classes (*ou clusters*).

Les techniques de clustering visent à ce que les individus d'une même classe partagent un degré élevé de similarité (*maximisation de la similarité intra-classe*) et que les différentes

1. Faire attention aux faux amis français/anglais : classification / clustering (non-supervisée). Quand on parle de "classification" en français, cela est traduit par "classification" en anglais. Par contre, le mot "classification" (non supervisée) est traduit par clustering en anglais.

classes soient aussi séparées que possible (*minimisation de la similarité inter-classe*). La similarité des individus est en général mesurée en terme de la distance géométrique entre les individus. Ces distances géométriques sont définies en *annexe C*. Et le nombre de classes k n'est pas forcément défini a priori, il peut être introduit par l'expert du domaine qui va déterminer l'intérêt et la signification des classes ainsi constituées.

Classification non supervisée / Clustering : Il s'agit d'identifier les populations d'un ensemble de données. On suppose qu'on dispose d'un ensemble d'individus que l'on note par $E = \{e_1; e_2; \cdots; e_r\}$ *caractérisé par un ensemble de descripteurs* Dr*. L'objectif du clustering est de structurer les données en classes homogènes, de façon à ce que les individus* e *d'une même classe soient les plus similaires possibles, qu'on note par* $C = \{C_1; C_2; \cdots; C_k\}$ *[CHY96], [JMF99].*

Le problème du clustering a été étudié dans plusieurs domaines, tels que l'analyse des données [CR93], [CS96], les bases de données spatiales [EKX95], [CA11], ou encore l'ECD [ZRL96], [AGGR98], [JBC13] notre domaine d'étude.

Il existe une très large famille de méthodes dédiées à la classification non supervisée. Dans ce chapitre, nous nous intéressons aux méthodes classiques : les méthodes de classification hiérarchique et les méthodes de partitionnement. Nous explicitons brièvement ces deux familles dans ce qui suit.

3.2.3 Revue des méthodes de classification

Vu le nombre de techniques de classification non supervisée parues dans la littérature, souvent à fort parfum heuristique, nous pouvons aujourd'hui les regrouper en deux grandes familles : les méthodes de classification qui suivent une stratégie hiérarchique et les méthodes qui suivent une stratégie par partitionnement. Pour une bonne introduction à ce domaine, nous pouvons consulter par exemple [Cel89] ou [KR90], [JK13].

Méthodes hiérarchiques Les méthodes hiérarchiques [CDG+89] cherchent à produire à partir d'une population donnée, une séquence de partitions imbriquées les unes dans les autres. Une telle suite de partition est souvent représentée par le biais d'un dendrogramme (*diagramme de Hasse arborescent dont les noeuds internes - matérialisés sous forme de segments horizontaux- sont alignées par niveau*) où chaque partition correspond à un niveau.

Nous citons parmi les méthodes hiérarchiques :

– les méthodes **ascendantes** (*agglomératives*) où les groupes sont fusionnés ;

– les méthodes **descendantes** (*divisives*) où un ou plusieurs groupes sont éclatés à chaque étape.

Les méthodes descendantes nécessitent l'usage de techniques de partitionnement à chaque étape de division des partitions, elles sont donc des méthodes non déterministes. Contrairement aux méthodes ascendantes, qui sont déterministes, et qui utilisent dans l'étape d'agrégation la notion d'indice d'agrégation (*voir annexe C*) permettant de mesurer la distance entre deux sous-ensembles d'individus. D'où, nous optons pour cette dernière méthode.

L'avantage des méthodes hiérarchiques par rapport aux méthodes de partitionnement, que nous définissons dans ce qui suit, réside dans le fait qu'il n'est pas nécessaire de définir a priori le nombre de classes. Elles visent à identifier les partitions en coupant le dendrogramme à un niveau satisfaisant. Chaque niveau correspond à une valeur numérique précise de l'indice d'agrégation. Plus la valeur de cet indice est élevée, plus les partitions (*ou classes d'individus*) sont hétérogènes.

Méthodes de partitionnement Les méthodes de partitionnement cherchent à diviser, de manière optimale, la population initiale en un nombre de classes fixé a priori. Il s'agit d'améliorer itérativement une partition initiale arbitraire en k classes, jusqu'à la convergence du critère choisi a priori.

Il existe dans la littérature statistique une profusion de méthodes et de critères de classification non hiérarchique, parmi lesquelles :

– **analyse en composantes principales** [Pea01] : cette méthode consiste à transformer des variables "corrélées" en nouvelles variables décorrélées les unes des autres ;
– **k-moyennes** [Mac67] : cette méthode est encore appelée algorithme des centres mobiles [Ben73], où chaque classe est représentée par son centre de gravité ;
– **nuées dynamiques** [Did71] : c'est une généralisation de l'algorithme des *k-moyennes*, où chaque classe est représentée par un sous-ensemble de la classe, appelé noyau ;
– **k-médoïdes** ou PAM (*Partitioning Around Medoids*) [KR90] : où chaque classe est représentée par un objet ou un représentant de cette classe (*medoïd*) ;
– **carte auto-organisatrice** (*Self Organizing Map*) [KSH01] : représente les données sur une carte 2D où chaque point de la carte représente un ensemble d'individus.

Dans la pratique, nous utilisons souvent les méthodes non-supervisées classiques type *k-moyennes* ou SOM. Ce dernier type de classification par cartes de Kohonen, regroupe les individus en fonction de leur distance du "centre" des différentes classes (*du type k-moyennes*), mais projette les données sur une grille de faible dimension. Il est donc peu performant pour la détection d'un petit nombre de classes. Dès lors, nous choisissons d'appliquer la méthode des

k-moyennes, qui par la pluralité de ses solutions (*puisque les partitions initiales sont arbitraires*), nous aidera d'avantage à identifier les classes fortes.

Nous définissons dans ce qui suit les méthodes choisies.

3.2.4 Choix de la procédure de classification

Dans ce qui suit, nous décrivons brièvement les méthodes de classification que nous allons utiliser, qui sont la classification ascendante hiérarchique et la technique des *k-moyennes*. La première méthode aura notre préférence pour son caractère déterministe. La seconde permettra, à partir de la contingence de ses résultats, de réduire l'incertitude de la validité de la classification obtenue par la recherche de formes stables, les formes fortes.

Nous commençons par définir la méthode hiérarchique.

Classification ascendante hiérarchique, ou CAH (*ou "par agrégation" ou "CAH"*) Elle consiste à regrouper itérativement les individus, en commençant par le bas (*les deux plus proches au sens de la distance deux à deux, telle que la distance euclidienne*) jusqu'à regrouper finalement tous les individus en une seule classe. À chaque étape, les deux clusters qui vont fusionner sont ceux dont la "*distance*" est la plus faible. L'utilisateur de cette méthode doit ainsi choisir un lien d'agrégation pour évaluer la "distance" entre deux groupes. Il existe plusieurs liens d'agrégation de Ward [War63], [Tuf05], que nous avons décrite dans l'*annexe C*. Le nombre de classes peut-être déterminé a posteriori, à la vue du dendrogramme.

Dans *l'annexe C*, nous définissons les critères utilisés et donnons plus de détails sur le principe de l'algorithme de la CAH.

Classification par les *k-moyennes* Cette méthode, encore appelée *algorithme des centres mobiles* [Ben73] est simple à appliquer et compréhensible. Elle représente l'une des méthodes de partitionnement les plus réputées [Mac67], [CDG+89]. Son but est de diviser l'ensemble des individus en k partitions (*clusters*) dans lesquelles chaque individu appartient à la classe dont le centre de gravité est le plus proche au sens de la distance choisie (*en général, euclidienne pour cette méthode*). Il s'agit de fixer à l'avance le nombre de classes k où un tirage aléatoire est effectué pour initialiser les k centres de ces classes. Dans une deuxième étape, une mise à jour des centres de chaque classe est effectuée. Le procédé (*affectation de chaque individu à un centre, détermination des centres*) est itéré jusqu'à convergence vers un optimum (*local*) ou un nombre d'itérations maximum fixé. Le principe de l'algorithme des *k-moyennes* est illustré dans l'*annexe C*.

L'inconvénient de la méthode des *k-moyennes* est qu'elle ne permet ni de découvrir le

nombre cohérent de classes, ni de visualiser la proximité entre les classes ou les objets. Ainsi, nous pouvons dire que la méthodes des *k-moyennes* et celle de la *CAH* sont complémentaires.

Ayant défini brièvement les méthodes de classification choisies, nous présentons dans ce qui suit les caractéristiques ainsi que la procédure de travail suivi pour la catégorisation des mesures.

3.2.5 Mise en oeuvre de la classification

Nous résumons dans ce qui suit les caractéristiques de la procédure de classification utilisée. Certaines justifications et définitions ont déjà été présentées ou le seront directement à la suite, et les autres sont données dans les *annexes B et C*. Nous présentons également notre procédure de travail pour la classification des mesures d'intérêt.

Caractéristiques de la classification Des choix de critères variés sont laissés à notre initiative, tels que :

- le choix des individus et de leurs variables : ***61 individus, 19 propriétés*** (*variables*).
- le format des variables : ***variables qualitatives nominales*** (*à l'issue d'une évaluation des mesures selon les propriétés, chapitre 2*) ;
- Choix de la mesure d'éloignement (*dissimilarité, distance*) entre individus : ***distance euclidienne*** ;
- Choix du critère d'homogénéité des classes à optimiser (*généralement inertie*) : ***critère de Ward*** [War63] ;
- Choix de la méthode de classification : ***classification ascendante hiérarchique suivie des k-moyennes*** (*centre mobile*) ;
- Choix du nombre de classes : nous utilisons les deux techniques de classification (*CAH et k-moyennes*) conjointement.
  L'idée est de remédier à l'inconvénient principal de la méthode des *k-moyennes*, qui est la saisie préalable du nombre de classes. Nous procédons alors dans un premier temps à la classification ascendante hiérarchique, où il s'agit d'identifier le nombre de classes par une coupure du dendrogramme généré à un niveau acceptable pour l'utilisateur. Par la suite, nous procédons à la méthode des *k-moyennes* et par l'introduction du nombre de classes obtenu à l'issue d'une CAH pour avoir une meilleure qualité de la classification.

Puisque nous procédons à la classification des mesures en utilisant des variables qualitatives, il serait donc usuel d'appliquer la distance euclidienne pour mesurer la dissemblance

entre les individus. Néanmoins, il serait aussi intéressant de tester avec la distance de Manhattan et de comparer les deux résultats. La dissemblance entre deux sous-ensembles disjoints est calculée à partir du critère d'agrégation de Ward de perte d'inertie minimum. Notre choix a porté sur ce critère parce qu'il est capable d'optimiser l'homogénéité des classes. En outre, il est basé sur un objectif clairement identifié de minimisation de l'inertie intra-classe, rendant ainsi l'algorithme et les résultats interprétables.

Suite à la sélection des critères de classification, nous pouvons maintenant présenter notre méthodologie de travail.

Processus d'analyse La procédure d'analyse du comportement des mesures d'intérêt est résumée dans ce qui suit :

Étape 1 : Préparation des données

Cette première étape consiste à traiter les données d'entrée, recueillies dans le cadre de l'étude des mesures d'intérêt selon les propriétés, pour les adapter aux méthodes de classification non supervisées que nous allons utiliser. Ayant des données qualitatives nominales, nous appliquons la technique la plus connue et la plus simple en analyse données, le codage disjonctif complet ou le codage $0/1$ sur la matrice de mesures-propriétés. Le but est de transformer les variables originelles en variables binaires. De ce fait, nous obtenons une nouvelle matrice avec des données binaires, permettant d'étudier toutes les caractéristiques d'une mesure et où aucune information n'est perdue. Le principe étant que chaque variable (*propriété*) prend la valeur 1 lorsque la modalité (*introduite en annexe B*) est présente pour un individu (*mesure d'intérêt*) , 0 sinon.

Étape 2 : Application des méthodes non supervisées

L'étape 2 consiste à appliquer les deux méthodes de classification non supervisée sur la matrice binaire obtenue lors de l'*étape 1* : la méthode de CAH dans un premier temps, puis la méthode des *k-moyennes*. Ces méthodes utilisent les critères choisis précédemment (*distance euclidienne, critère de ward*) pour donner une nouvelle structure aux données binaires sous forme de catégories de mesures. Chaque méthode propose sa propre classification.

Étape 3 : Classification consensuelle des mesures d'intérêt

L'étape 3 consiste à trouver un consensus sur la classification des mesures en confrontant les résultats obtenus par les deux méthodes dans l'étape précédente.

Étape 4 : Interprétation

L'étape 4 consiste à interpréter les catégories de mesures identifiées afin de comprendre le comportement des mesures d'une même classe.

Étape 5 : Comparaison

Finalement, l'étape 5 consiste à comparer les résultats de la classification consensuelle,

obtenue dans l'*étape 3*, avec ceux des travaux existant dans la littérature.

Nous procédons dans ce qui suit à la classification des mesures d'intérêt en suivant les 3 premières étapes de notre méthodologie de travail.

3.3 Classification des mesures d'intérêt

Cette section s'intéresse à la classification des mesures d'intérêt. Nous analysons dans un premier temps les données d'entrée par la recherche d'une redondance au niveau des lignes/colonnes de la matrice d'évaluation des mesures (*décrite dans le chapitre 2, pages 70 et 71*). Par la suite, nous appliquons les deux méthodes de classification non supervisée *CAH* et *k-moyennes*.

3.3.1 Les données d'entrée

Nous reprenons la matrice d'évaluation des mesures selon les propriétés, qui représente nos données d'entrée, afin de s'assurer qu'il n'y a pas une possibilité de la restreindre. Dans le cas où nous identifions des groupes de mesures aux comportements identiques ou de propriétés redondantes, nous cherchons à réduire les dimensions de cette matrice par la suppression de la redondance. Pour ce faire, nous recherchons tout d'abord l'ensemble des mesures dont les valeurs pour chacune des 19 propriétés sont identiques. Nous recensons les 7 groupes suivants :

- G_{s1} = {*Coefficient de corrélation, Nouveauté*},
- G_{s2} = {*Confiance causale, Confiance-confirmée causale, Fiabilité négative*},
- G_{s3} = {*Cosinus, Czekanowski-Dice*},
- G_{s4} = {*Dépendance causale, Leverage, Spécificité*},
- G_{s5} = {*Force collective, Ratio des chances*},
- G_{s6} = {*Gini, Information mutuelle*},
- G_{s7} = {*Jaccard, Kulczynski*}.

Suite à la détection de ces 7 groupes de mesures $\{G_{s1}, \cdots, G_{s7}\}$ et afin d'éviter de la redondance dans nos données, nous ne gardons qu'une seule mesure par groupe. Nous sommes désormais en présence d'une matrice de 52 mesures puisque nous ne retenons que 7 mesures parmi les 16 citées ci-dessus (*une de chaque groupe*).

Par la suite, nous vérifions si des propriétés ne sont pas redondantes. Pour cela, nous recherchons si une propriété possède des valeurs identiques pour chacune des 52 lignes avec une autre propriété. Nous n'avons identifié aucune relation de ce type, ce qui nous révèle qu'il n'y a pas de propriétés identiques.

Par conséquent, nous sommes à présent avec une matrice de 52 mesures et 19 propriétés, construite à partir de variables qualitatives nominales (*une définition de ces types de variables est présente en annexe B*). Nous appliquons un codage disjonctif complet sur l'ensemble de ces variables qui nous conduit à l'obtention de 39 variables binaires. Une nouvelle structure de notre matrice prend donc forme, avec 52 *mesures* $\times$ 39 *variables binaires*.

L'étape de préparation des données étant achevée, nous allons maintenant procéder à la deuxième étape de notre processus d'analyse et appliquer les algorithmes de classification sur la nouvelle matrice.

3.3.2 Classification obtenue par une méthode de CAH

Nous effectuons une classification ascendante hiérarchique (*CAH*) de *52* mesures en partant du tableau disjonctif complet. Pour ce faire, nous utilisons le logiciel Matlab, la distance euclidienne entre paires de mesures et la distance de Ward pour la phase d'agrégation. La *figure 3.1* restitue cette classification pour la distance de Ward. Comme la perte d'inertie interclasse doit être la plus faible possible, nous coupons le dendrogramme à un niveau où la hauteur des branches est élevée, ce qui correspond aux branches colorées du dendrogramme.

En choisissant la distance de Manhattan, nous obtenons des résultats sensiblement similaires, comme le montre aussi Vaillant dans sa thèse [Vai06].

Cette classification nous révèle les 8 groupes de mesures suivants :

- G_{c1} = {*Indice de vraisemblance du lien (IVL), Intensité d'implication (II)*}
- G_{c2} = {*IIER, IIE, IPD, IP3E, IPEE*}
- G_{c3} = {*Variation du support à double sens, Pearl*}
- G_{c4} = {*Indice d'implication, Fukuda, Gini, J-mesure, Dépendance, Dépendance pondérée, Prévalence, Couverture*}
- G_{c5} = {*VT100, Précision, Jaccard, Support, Cosinus, Rappel, Dépendance causale, Confirmation causale, Confiance causale*}
- G_{c6} = {*Sebag, Moindre contradiction, Confirmation descriptive, Taux d'exemples, Ganascia, Laplace, Confiance*}
- G_{c7} = {*Zhang, M_{GK}, Y de Yule, Q de Yule, Goodman, Piatetsky-Shapiro, Coefficient de corrélation*}

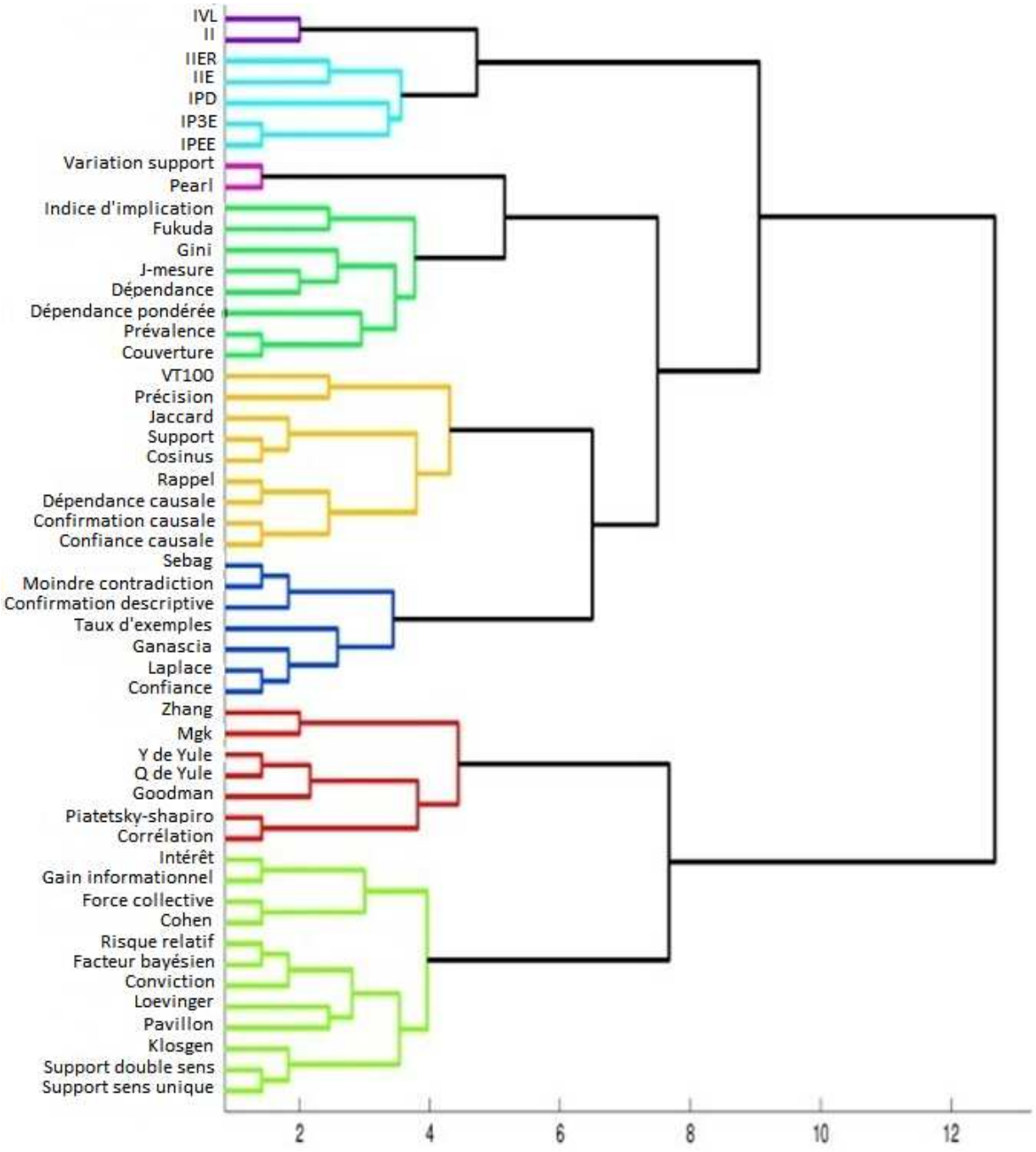

FIGURE *3.1: Classification ascendante hiérarchique utilisant le critère de Ward.*

– G_{c8} = {*Intérêt, Gain informationnel, Force collective, Cohen, Risque relatif, Facteur bayésien, Conviction, Facteur de certitude, Pavillon, Klosgen, Support à double sens, Support à sens unique*}

Les résultats de la classification révélée par la méthode hiérarchique présentent une partition unique pour un seul niveau de coupure du dendrogramme.

Après avoir effectué cette première classification des mesures, nous allons confronter ces résultats avec une deuxième technique : une version de la méthode des *k-moyennes* obtenue par le logiciel *Matlab*.

3.3.3 Classification obtenue par une version des k-moyennes

Nous réalisons maintenant un partitionnement de 52 mesures au moyen de la méthode des *k-moyennes*. Ainsi, de même que pour la classification hiérarchique, nous utilisons le logiciel *Matlab* et nous retenons les mêmes données d'entrée. Cette méthode permet à l'utilisateur de spécifier la mesure de distance qu'il faut utiliser dans le processus de minimisation. Nous gardons donc la distance euclidienne. Pour d'amples détails sur la description de cette fonction, vous pouvez consulter la fonction en ligne [2].

Toutefois, et pour le bon fonctionnement de la technique des *k-moyennes*, un autre critère doit être introduit, qui est le nombre de classes désiré. Dès lors, nous choisissons 8 au vu des résultats de la *CAH* et nous obtenons le partitionnement présenté ci-après. Au fur et à mesure que nous présentons ces 8 nouvelles classes identifiées, nous discutons de la cohérence des résultats obtenus avec la première technique.

– G_{p1} = {*indice de vraisemblance du lien, intensité d'implication, IIER*}
  Ce groupe est très proche du groupe Gc1 puisque nous avons $G_{p1} = G_{c1} \cup$ {*IIER*}.
– G_{p2} = {*IIE, IPD, IP3E, IPEE*}
  Ce groupe est très proche du groupe G_{c2} puisque nous avons $G_{p2} = G_{c2}$ - {*IIER*}. Nous avons l'égalité suivante : $G_{p1} \cup G_{p2} = G_{c1} \cup G_{c2}$, ce qui montre une certaine cohérence dans les résultats obtenus puisque nous sommes en présence de tous les indices de la famille de la vraisemblance du lien.
– G_{p3} = {*Variation du support, Pearl, Indice d'implication, Gini, J-mesure, Dépendance, Prévalence, Couverture*}
  Ce groupe est proche du groupe G_{c4} puisque nous avons :
  $G_{p3} = G_{c3} \cup G_{c4} \cup$ {*Fukuda, Dépendance pondérée*}. Il est à noter que le groupe G_{c3},

2. http ://www.mathworks.fr/fr/help/stats/kmeans.html

composé des mesures *Variation du support* et *Pearl*, est le groupe le plus proche du groupe G_{c4} (*voir le dendrogramme de la figure 3.1*).

- G_{p4} = {*Précision, Jaccard, Support, Cosinus, Rappel, Dépendance causale, Confirmation causale, Confiance causale, Fukuda, Dépendance pondérée*}
  Ce groupe est similaire au groupe G_{c5} puisque nous avons :
  $G_{p4} = G_{c5} \cup$ {*Fukuda, Dépendance pondérée*} - {*VT100}*.
- G_{p5} = {*Sebag, Moindre contradiction, Confirmation descriptive, Taux d'exemples, Ganascia, Laplace, Confiance*}
  Ce groupe est identique au groupe G_{c6}.
- G_{p6} = {*Zhang, M_{GK}, Y de Yule, Q de Yule, Goodman*}
  Ce groupe est similaire au groupe Gc_7 puisque nous avons :
  $Gc_7 = Gp_6 \cup$ {*Piatetsky-Shapiro, Coefficient de corrélation*}.
- G_{p7} = {*Intérêt, Gain informationnel, Risque relatif, Facteur bayésien, Conviction, Facteur de certitude, Pavillon, Klosgen, Support à double sens, Support à sens unique*}
  Le groupe G_{p7} est très proche du groupe G_{c8} puisque nous avons 10 mesures en commun sur 12. Nous avons l'égalité suivante : $G_{c8} = G_{p7} \cup$ {*Force collective, Cohen*}.
- G_{p8} = {*VT100, Piatetsky-Shapiro, Coefficient de corrélation, Force collective, Cohen*}
  Contrairement aux autres groupes G_{pi} ($i = \{1, \cdots, 7\}$), ce groupe n'est similaire à aucun des groupes G_{cj} (*j = {1, .., 8}*) puisque ces 5 mesures sont issues des groupes G_{c5}, G_{c7} et G_{c8}.

Les 8 groupes de mesures ainsi obtenus résultent d'une version des *k-moyennes* puisque la méthode appliquée est une heuristique qui n'est pas déterministe. Elle présente toutefois une autre particularité, autre que la fixation préalable du nombre de classes, la contingence de ses résultats. Cette méthode de partitionnement ne produit pas forcément une solution optimale mais plutôt un optimum local. Cette variabilité des résultats est due au choix initial des centres de classes qui se fait de manière arbitraire.

Face à la dispersion possible des résultats de la classification des mesures par les *k-moyennes*, nous cherchons dans ce qui suit à déterminer des formes stables, des formes fortes. Cette notion de forme forte étant une première réponse à ce problème puisqu'elle consiste à s'assurer de l'obtention de façon stable de classes ou portions de classes de mesures, quel que soit le choix des centres initiaux.

3.3.4 Classes fortes

Pour déterminer les classes fortes des mesures d'intérêt, nous allons répéter l'exécution de l'algorithme des *k-moyennes* une dizaine de fois et identifier les mesures qui se trouvent systé-

matiquement dans le même groupe quelle que soit la partition initiale [Did82]. L'intersection des résultats révélés par les *10* répétitions nous permet de discerner les classes fortes suivantes :

- G_{f1} = {*Indice de vraisemblance du lien (IVL), Intensité d'implication (II), IIER*}
- G_{f2} = {*IIE, IPD, IP3E, IPEE*}
- G_{f3} = {*Variation du support à double sens, Pearl*}
- G_{f4} = {*Indice d'implication, Dépendance, Couverture*}
- G_{f5} = {*Jaccard, Support, Cosinus, Rappel, Dépendance causale, Confirmation causale, Confiance causale*}
- G_{f6} = {*Sebag, Moindre contradiction, Confirmation descriptive, Taux d'exemples, Ganascia, Laplace, Confiance*}
- G_{f7} = {*Zhang, M_{GK}, Y de Yule, Q de Yule, Goodman*}
- G_{f8} = {*Intérêt, Gain informationnel, Risque relatif, Facteur bayésien, Conviction, Facteur de certitude, Pavillon, Klosgen, Support à double sens, Support à sens unique*}

Ces 8 classes fortes, ou classes stables sont alors issues des essais multiples de l'algorithme des *k-moyennes*. À partir de ces *10* essais, nous constatons que certaines mesures {*VT100, Piatetsky-Shapiro, Coefficient de corrélation, Force collective, Cohen, Dépendance pondérée, Fukuda, Prévalence, Précision, J-mesure et Gini*} n'étaient pas présentes parce qu'elles déviaient parfois de leurs groupes. Parmi ces mesures, nous remarquons que les cinq premières forment le groupe G_{p8}, identifié par l'algorithme des *k-moyennes* dans la *sous-section 3.3.3*. Les résultats obtenus dans cette *sous-section* peuvent être considérés comme étant une solution acceptable révélée par la méthode de partitionnement. Cette solution étant la plus courante après les nombreuses exécutions (*10 répétitions*) réalisées par l'algorithme des *k-moyennes* sans toutefois avoir de garantie d'optimalité globale de la partition retenue [Gué06].

Étant donné l'ensemble des résultats de la classification obtenus par la *CAH*, par une version des *k-moyennes*, également par les classes fortes, nous discuterons dans ce qui suit ces différentes solutions afin de dégager un consensus.

3.3.5 Classification définitive

Après avoir discuté dans la *sous-section 3.3.3* sur la cohérence des résultats obtenus par les deux techniques (*CAH et k-moyennes*) et identifié les classes fortes, nous dégageons un consensus sur la classification. La *figure 3.2* révèle ce consensus et nous restitue les classes C_1 à C_7 de mesures extraites communes aux deux techniques. Néanmoins, pour certaines mesures aucun consensus n'a été trouvé ce qui fait varier le nombre de classes de 8 à 7 classes

de mesures. Par exemple, le groupe G_{p8} identifié par la méthode des *k-moyennes* n'est pas vérifié par la méthode hiérarchique et donc il n'y a pas eu de consensus concernant ce groupe. Dans la *figure 3.2*, nous mentionnons ces mesures et donnons, lorsque c'est possible, les deux groupes (*ou classes*) d'appartenance de ces mesures. Nous étiquetons aussi les flèches par "*c*" et "*p*" pour indiquer quelle technique a rassemblé les mesures dans le groupe pointé (*c = classification hiérarchique ou p = partitionnement*). Les classes fortes étant celles représentées par les cadrans dont le fond est coloré. Pour finir, nous rappelons dans le cadran inférieur du milieu, les mesures identiques *i.e.* ayant une même définition (*ou formule*) mais portant des noms différents.

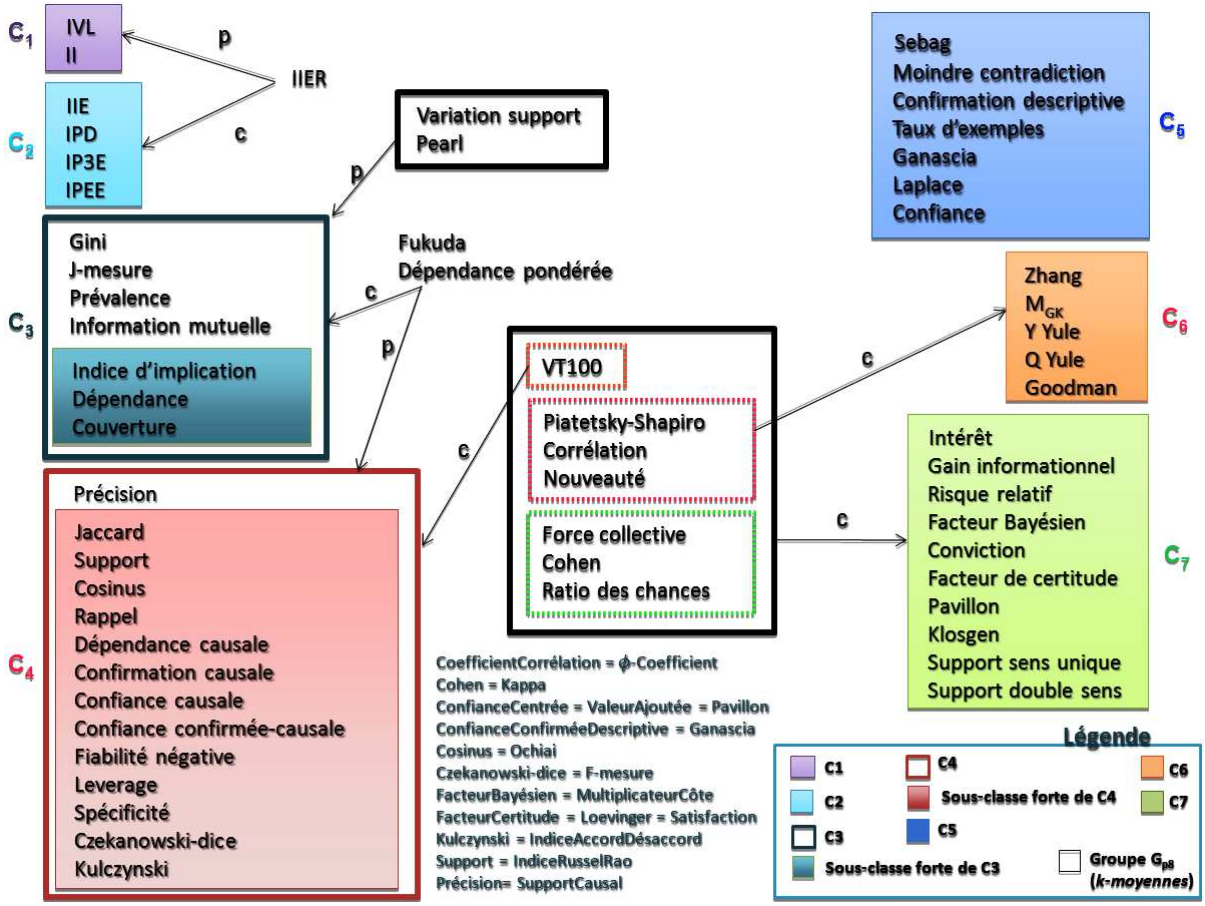

FIGURE *3.2: Groupes ou classes de mesures.*

Ayant schématisé le consensus sur la classification, nous passons dans la section suivante à la quatrième étape de notre processus l'analyse, qui est la recherche d'une sémantique aux 7 classes extraites

3.4 Étude des classes

Nous cherchons à partir de cette section à expliquer chacune des 7 classes $C_i(i = 1, .., 7)$ identifiées précédemment. Pour ce faire, nous résumons toutes les propriétés qu'elles vérifient dans la *table 3.1*. Nous rajoutons dans cette dernière un symbole par rapport à la matrice d'origine, le caractère "?", qui a la signification "*indéterminé*" c'est-à-dire que les mesures de la classe C_i prennent différentes valeurs pour la propriété $P_j(j = 1, .., 19)$ concernée. Dans le cas où la propriété est contredite une seule fois, nous indiquons la valeur de la propriété majoritaire. Ainsi "$0?$" signifie que toutes les mesures de la classe C_i sauf une seule mesure, prennent la valeur "0" pour la propriété P_j.

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Classes
?	1	1	1	1	1	1	0	0	1	1	2	?	0	0	0	1	1	0	C_1
1	1	1	1	0?	1?	0	0	?	0	0	2	0?	0	0	0	1	1	1?	C_2
1	?	0?	0	0	0	?	0	0	0?	0	?	0	0	0	?	0	0?	?	C_3
?	1	?	1?	?	1?	0	?	0	0	0	?	0	0	0	0?	0	0	1	C_4
1	1	?	1	0	0	0	?	1	0	0	?	0	0	?	0	0	0	1?	C_5
?	1	1	1	1	0	1	1	0	1	1	?	?	?	1	?	0	0	1	C_6
?	1	?	?	1	1?	1	0?	0	1	1	?	0?	0	0?	0	0	0	1	C_7

TABLE *3.1: Caractéristiques des sept classes détectées*

En résumant l'ensemble des propriétés vérifiées par chacune des *7* classes dans ce tableau, et en se référant au dendrogramme de la *figure 3.1* et à la *table 3.9* où apparaît une notion de proximité entre les mesures, nous pouvons préserver l'utilisateur de choisir des mesures trop similaires. Ce qui lui évitera de prendre des indices appartenant à la même classe. Également, nous pouvons discerner les classes qui se ressemblent le plus entre elles. Nous prenons comme exemple les classes C_4 et C_7 qui diffèrent principalement sur l'évaluation de 4 propriétés uniquement.

Quant à la recherche d'une sémantique, nous présentons dans ce qui suit une interprétation approfondie des différentes classes. Ce tableau synthétique *3.1* ainsi que la matrice de distance (*table 3.9*) peuvent être le support à cette interprétation comme nous allons l'illustrer pour les classes C_1 à C_7 obtenues. Pour ce faire, nous suivons la démarche suivante, constituée de 3 étapes essentielles :

Étape a : *trouver des relations mathématiques ou une interprétation sémantique* par une phrase entre les mesures d'un même groupe ;

Étape b : *étude du comportement des mesures selon les propriétés formelles*. Il s'agit dans un premier temps d'identifier les caractéristiques communes aux mesures d'un même groupe. Ensuite, nous allons étudier la proximité existant entre ces mesures au moyen de la matrice de distance illustrée dans la *table 3.9*. Pour cela, nous jugeons que deux mesures sont proches si

elles sont distantes d'une valeur qui soit inférieure à $2,74$. Cette valeur est égale à la moyenne de la distance minimale et maximale identifiées par la *table 3.9* et qui valent respectivement $0,00$ et $5,48$.

Durant cette étape et dans le cas où cela semble nécessaire pour l'interprétation, nous aurons besoin de tracer des courbes d'évolution des mesures en fonction du nombre d'exemples. Pour ce faire, nous retenons les paramètres illustrés dans la figure 1.4 : la taille de l'ensemble prémisse est de *174*, la taille de l'ensemble conclusion est de *400* et pour finir la taille de l'ensemble des données est de *600* ($n_X = 174$, $n_Y = 400$ *et* $n = 600$). Nous aurions pu choisir des tailles différentes pour ces différents ensembles et aurions obtenu des courbes similaires avec la contrainte suivante respectée : $n_X \leq n_Y \leq n$;

Étape c : *appliquer une classification ascendante hiérarchique*. Cette étape du processus ne sera appliquée qu'en cas de besoin, comme par exemple si le nombre de mesures d'un même groupe est assez élevé. Son but est d'obtenir des sous-groupes facilitant l'interprétation de la classe en question.

Finalement et pour chaque classe, nous proposons une ou plusieurs mesures référentes, capables de représenter les mesures de leur classe. Le choix de la (les) mesure(s) référente(s) se fait suite à la consultation de la table de distance (*table 3.10*) entre chaque mesure et le centre de gravité. Les mesures les plus proches sont les mesures référentes.

Dans ce qui suit, nous appliquons la démarche précédemment décrites sur les 7 classes de mesures que nous avons obtenues.

3.4.1 Étude des classes C_1 et C_2

C_1 et C_2 demeurent les deux classes les plus faciles à interpréter puisque nous retrouvons tous les indices de la famille de l'indice de vraisemblance du lien [Ler70], indice fondateur.

3.4.1.1 Étude de la classe C_1

La classe C_1 possède les indices d'origine : l'indice de vraisemblance du lien (*IVL*) et l'intensité d'implication (*II*).

- *Étape a* : Les deux mesures de cette classe sont très proches puisque l'indice de vraisemblance du lien recherche si le nombre d'exemples est significativement élevé alors que l'intensité d'implication évalue si le nombre de contre-exemples (*ceux qui vérifient la prémisse mais qui ne vérifient pas la conclusion*) est significativement faible.
- *Étape b* : La *table 3.2* rappelle les caractéristiques de la classe C_1 par la présentation des propriétés vérifiées par ce couple de mesures. Nous remarquons une grande cohérence dans l'évaluation des propriétés pour ces deux mesures. Seules les propriétés P_3 et P_{15}

sont évaluées différemment. En outre, ces 2 mesures sont distantes d'une valeur égale à $2,00$ selon la *table 3.9*. En considérant la distance moyenne égale à $2,74$, nous jugeons que ces deux mesures sont assez proches.

– *Étape c* : cette étape n'est pas nécessaire.

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Classes
1	1	1	1	1	1	1	0	0	1	1	**2**	1	0	0	0	1	1	0	II
0	1	1	1	1	1	1	0	0	1	1	**2**	0	0	0	0	1	1	0	IVL
?	1	1	1	1	1	1	0	0	1	1	2	?	0	0	0	1	1	0	Classe C_1

TABLE *3.2: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 1.*

3.4.1.2 Étude de la classe C_2

Pour la classe C_2, nous retrouvons les mesures d'intensité d'implication entropiques (*IIE, IP3E)* avec l'indice probabiliste d'écart à l'équilibre (*IPEE*) ainsi que l'indice probabiliste discriminant (*IPD*).

– *Étape a* : Ces mesures sont issues d'une idée commune : évaluer la significativité d'un nombre (*nombre d'exemples ou de contre-exemples*), en le combinant pour certaines mesures (*IIER, IIE, IP3E*) avec un indice entropique afin que la mesure soit discriminante dans le cas de données volumineuses. Quant à l'IPD, cet indice normalise l'intensité d'implication afin que celle-ci soit discriminante dans le cas de données volumineuses en évaluant une règle par rapport à l'ensemble des règles valides.

– *Étape b* : Dans la *table 3.3*, nous restituons les propriétés vérifiées par les mesures de la classe C_2 et qui sont les suivantes : P_3, P_4, P_5, P_6, $\overline{P_9}$, $\overline{P_{10}}$, $\overline{P_{12}}$, $\overline{P_{13}}$, P_{14}, $\overline{P_{16}}$, $\overline{P_{17}}$, $\overline{P_{18}}$, P_{19} et P_{20}. Outre les propriétés vérifiées par presque toutes les mesures à l'exception d'une seule et sont : $\overline{P_7}$, P_8, $\overline{P_{15}}$ et P_{21}, les mesures de C_2 sont donc toutes des mesures statistiques, non symétriques, qui croissent en fonction du nombre d'exemples et possédant des valeurs variables au niveau de l'indépendance et de l'implication logique. Selon la *table 3.9*, les deux mesures les plus proches de cette classe sont *IPEE* et *IP3E* pour une distance égale à $1,41$, et celles qui sont les plus éloignées sont *IPEE* et *IPD*, avec $d = 3,16$.

– *Étape c* : cette étape n'est pas nécessaire.

Noyau ou mesure référente de C_1 et C_2 Nous cherchons maintenant à aider d'avantage l'utilisateur et lui proposer une mesure référente pour la classe C_2. Pour ce faire, nous étudions les distances séparant les mesures avec le centre de gravité. Ces distances sont obtenues suite à l'exécution de l'algorithme des *k-moyennes* pour l'identification des 8 groupes de mesures de

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Classes
1	1	1	1	1	0	0	0	0	0	0	**2**	0	0	0	0	1	1	1	IIE
1	1	1	1	0	1	0	0	0	0	0	**2**	1	0	0	0	1	1	1	IPD
1	1	1	1	0	0	0	0	1	0	0	**2**	0	0	0	0	1	1	1	IP3E
1	1	1	1	0	0	0	0	1	0	0	**2**	0	0	0	0	1	1	0	IPEE
1	1	1	1	0 ?	1 ?	0	0	?	0	0	2	0 ?	0	0	0	1	1	1 ?	Classe C_2

TABLE *3.3: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 2.*

la *sous-section 3.3.3*. La *table 3.10* illustre ces valeurs de distance des *52* mesures étudiées avec les centres de ces groupes. La mesure la plus proche du centre peut représenter sa classe.

Ainsi, la classe C_2 peut être représentée par la mesure *IP3E* puisqu'elle est la plus proche de son centre (*d=1,13*), par rapport aux trois autres mesures de C_2 (*table 3.10*). Bien évidemment, le noyau peut aussi être formé de plus qu'une mesure.

Pour la classe C_1, l'utilisateur peut choisir l'une des deux mesures *Intensité d'implication* ou *IVL* puisque le centre de gravité est équidistant des deux mesures.

3.4.2 Étude de la classe C_3

La classe C_3 comprend les 7 mesures suivantes : {*Indice d'implication*, *Dépendance*, *Gini*, *J-mesure*, *Prévalence*, *Couverture* et *Information mutuelle*}.

- *Étape a* : L'interprétation de cette classe semble être difficile, puisque non seulement nous sommes incapables de les expliquer par une phrase, mais aussi la *table 2.5, page 73* ne révèle aucune relation ou dépendance entre les différentes mesures de C_3.
- *Étape b* : La *table 3.4* rappelle les caractéristiques de cette classe et exhibe les propriétés non vérifiées par ses mesures. Elle nous fait remarquer que l'ensemble des mesures de C_3 ne partagent que des propriétés négatives[3] (P_3, $\overline{P_6}$, $\overline{P_7}$, $\overline{P_8}$, $\overline{P_{10}}$, $\overline{P_{11}}$, $\overline{P_{13}}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{17}}$, $\overline{P_{19}}$) à l'exception de la P_3. Ainsi, le fait d'appliquer le codage disjonctif complet peut entraîner la construction de cette classe puisque les "0" identifiés deviendront des "1" et par conséquent les mesures de C_3 vérifieront des propriétés non souhaitables permettant de les regrouper ensemble.
- *Étape c* : Nous abordons cette étape de notre processus afin de voir si des sous-groupes de cette classe peuvent être interprétés. Pour ce faire, nous appliquons la méthode hiérarchique *CAH* sur un extrait du tableau disjonctif complet correspondant à la classe C_3. La *figure 3.3* présente le dendrogramme approprié à cette classe et met en exergue 3 sous-groupes de mesures :

  1. le premier sous-groupe ne comprend qu'une seule mesure, "*Indice d'implication*" ;

3. Les propriétés négatives sont les négations des propriétés formelles décrites dans le chapitre 2, *e.g.* la négation de la propriété P_3 est $\overline{P_3}$ qui désigne une mesure symétrique

2. le deuxième sous-groupe contient les mesures *Dépendance*, *Gini* et *J-mesure* ;
3. le dernier sous-groupe comprend les mesures restantes *Prévalence* et *Couverture*.

La matrice de distance entre les mesures, illustrée dans la *table 3.9* permet de nous renseigner d'avantage sur cette classe par la découverte de mesures qui sont proches entre elles et celles qui sont plutôt éloignées. Ainsi, en s'appuyant sur cette matrice, sur le dendrogramme de la *figure 3.3* et sur la matrice de la *table 3.10*, nous interprétons les 3 sous-groupes. Nous commençons par le premier qui comprend uniquement la mesure *Indice d'implication* (*IndImp*). Nous remarquons que cette dernière est la mesure la plus distante du centre de C_3 ($d = 4,81$), qui est donc placée à l'extrémité de cette classe, comme nous le voyons également sur le dendrogramme de la *figure 3.3*. En outre, la matrice de distance (*table 3.9*) montre que *l'Indice d'implication* est distant des autres mesures de C_3 selon les valeurs suivantes : *d(IndImp, Gini)* $= 3,46$ *; d(IndImp, Dép)* $= 2,45$ *; d(IndImp, Prév)* $= 2,83$ *; d(IndImp, Jmes)* $= 3,16$ *et d(IndImp, Couv)* $= 3,16$.
Selon la *table 3.4*, *l'Indice d'implication* est le seul, parmi les autres mesures de C_3, à posséder une allure linéaire ($P_{14} = 1$), qui croît en fonction du nombre d'exemples.
Le deuxième sous-groupe comprend trois mesures *Dépendance*, *Gini* et *J-mesure* qui selon *figure 3.4*, possèdent une allure concave et se croisent à l'indépendance. Nous remarquons à partir de la matrice de distance que la mesure *Gini* est proche des deux autres mesures pour une même distance *d(Gini, Dép)= d(Gini, Jmes)= 2,45*.
Finalement, le troisième sous-groupe qui contient deux mesures *Prévalence* et *Couverture*. Ces dernières sont linéaires et parallèles, comme le montre la *figure 3.4*, insensibles au nombre d'exemples. En termes de distance, nous remarquons que ces deux mesures (*Prévalence et Couverture*) sont les plus proches au niveau de la classe C_3, avec une distance *d(Prév, Couv) = 1,41*. De plus, la *table 3.4* montre que seule la propriété P_4 n'est pas vérifiée par ces deux mesures.

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	**Classes**
1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	Indice d'implication
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	Gini
1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	J-Mesure
1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	Dépendance
1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Prévalence
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Couverture
1	?	0?	0	0	0	?	0	0	0?	0	?	0	0	0	?	0	0?	?	Classe C_3

TABLE *3.4: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 3.*

Noyau ou mesure référente de C_3 Nous suivons la même méthode que pour la classe C_2, nous cherchons entre les 6 mesures de la classe C_3, celles qui possèdent une distance mini-

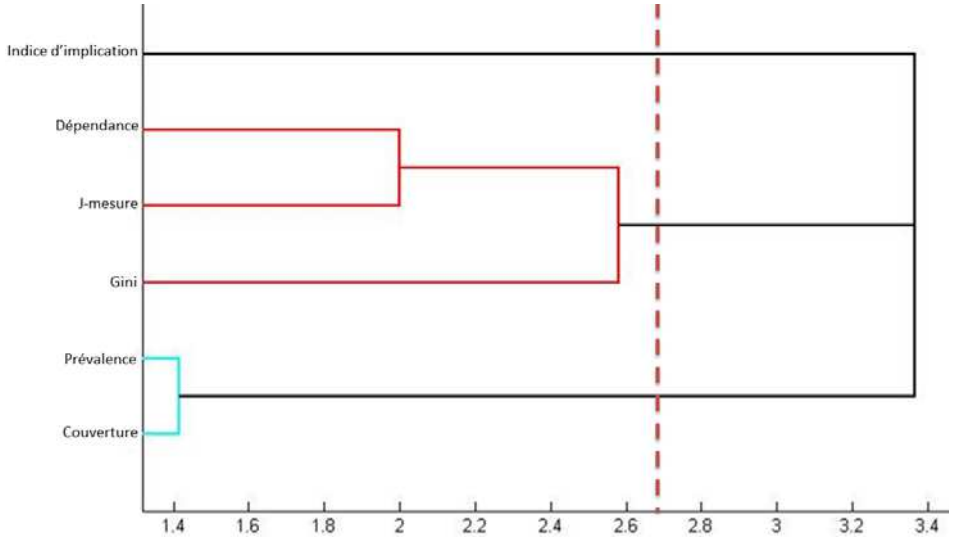

FIGURE *3.3: Classification hiérarchique des mesures de la classe* C_3.

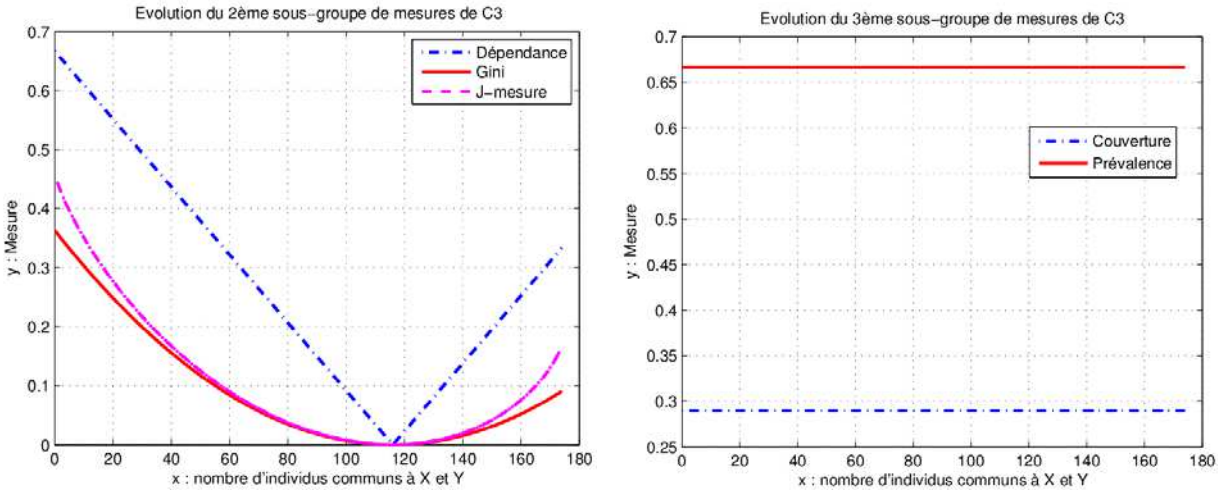

FIGURE *3.4: Évolution du 2ème sous-groupe (gauche) et du 3ème sous-groupe (droite) de mesures de* C_3 *en fonction du nombre d'exemples.*

male avec le centre de gravité. Dès lors, nous consultons la *table 3.10* et nous découvrons que la mesure *Dépendance* est la plus proche du centre de gravité de C_3 avec une distance égale à $1,81$. Elle représente ainsi les mesures de sa classe. Cette mesure appartient au deuxième sous-groupe, il est donc le plus proche du centre, suivi du troisième sous-groupe qui est représenté par la mesure couverture pour $d = 3,31$ et enfin le premier sous-groupe qui contient une seule mesure, l'*Indice d'implication*.

Nous passons maintenant à l'étude de la classe C_4.

3.4.3 Étude de la classe C_4

La classe C_4 contient les indices suivants : *Précision, Jaccard, Support, Cosinus, Rappel, Dépendance causale, Confirmation causale, Confiance causale, Confiance-confirmée causale, Fiabilité négative, Leverage, Spécificité, Czekanowski-Dice, et Kulczynski*. L'expression de ces indices ainsi que leurs références bibliographiques sont données en *annexe A*.

– *Étape a* : Vu le nombre relativement important des mesures présentes dans cette classe (*c'est la classe dont la cardinalité est la plus élevée*), il est difficile de trouver une sémantique aussi précise que pour les classes C_1 et C_2. Nous pouvons cependant en donner une pour un ensemble plus restreint de mesures : *Jaccard*, *Support*, *Cosinus*, *Czekanowski-Dice*, *Kulczynski* et *Rappel*. Ces mesures possèdent toutes en numérateur $P(XY)$ uniquement, qui correspond à la mesure *Support*, et sont toutes symétriques à l'exception de la mesure *Rappel*. Nous rappelons les expressions de ces 5 mesures :

- $Jaccard = \frac{P(XY)}{P(X)+P(Y)-P(XY)} = \frac{P(XY)}{P(X\overline{Y})+P(Y)}$
- $Cosinus = \frac{P(XY)}{\sqrt{P(X)P(Y)}}$
- $Czekanowski - Dice = \frac{2\times P(XY)}{P(X)+P(Y)} = \frac{2\times P(XY)}{P(XY)+1-P(\overline{X}\overline{Y})}$
- $Kulczynski = \frac{P(XY)}{P(X\overline{Y})+P(\overline{X}Y)}$
- $Rappel = \frac{P(XY)}{P(Y)}$

À la vue de ces différentes expressions, nous pouvons donc en déduire que ces mesures, y compris le *support*, auront une valeur fixe égale à 0 dans le cas de l'incompatibilité ($P(XY) = 0$). Nous comprenons également la non croissance trouvée en fonction de la taille n de l'ensemble des données ($P_7 = 0$), chose prouvée par la *table 3.5* qui restitue les propriétés vérifiées par ces 5 mesures outre le *support*.

– *Étape b* : D'après la *table 3.1*, ces 14 mesures vérifient les 11 propriétés suivantes : P_4, $\overline{P_9}$, $\overline{P_{11}}$, $\overline{P_{12}}$, $\overline{P_{13}}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{17}}$, $\overline{P_{19}}$, $\overline{P_{20}}$, P_{21}. Outre les propriétés P_6, P_8 et $\overline{P_{18}}$, qui sont vérifiées par presque toutes les mesures à l'exception d'une seule.
Comme pour la classe précédente C_3, nous allons étudier l'évolution de ces différentes mesures en fonction du nombre d'exemples. La *figure 3.5* restitue cette évolution. Nous vérifions la valeur nulle prise par ces mesures dans le cas de l'incompatibilité. Nous obtenons deux types de courbes :

- une droite pour les mesures *Support*, *Cosinus*, *Czekanowski-Dice* et *Rappel*,
- une demi-parabole pour les mesures *Jaccard* et *Kulczynski*.

En terme de distances (*table 3.9*), nous remarquons que la mesure *Cosinus* est très proche des mesures *Jaccard, Support et Rappel* pour une même distance $d = 1,41$ la

séparant de chacune d'elles. Cette même valeur de distance rapproche d'une part la mesure *Dépendance causale* aux mesures *Rappel et Confirmation causale* et d'autre part la mesure *Confirmation causale* à la mesure *Confiance causale*.

– *Étape c* : Cette étape du processus s'avère importante vu le nombre de mesures appartenant à cette classe. Discerner des sous-groupes est peut-être intéressant dans la mesure où nous pouvons identifier des ensembles de mesures facilement interprétables. Ainsi, nous appliquons la méthode hiérarchique sur un extrait du tableau disjonctif complet correspondant à la classe C_4. Le dendrogramme de la *figure 3.6* met en exergue les 3 sous-groupes de mesures suivants :

1. {*Rappel, Dépendance causale*} ;
2. {*Jaccard, Support, Cosinus*} ;
3. {*Précision, Confiance causale, Confirmation causale*}.

Le premier sous-groupe réunit les deux mesures *Rappel* et *Dépendance causale*. Selon la *table 3.5*, ces dernières partagent toutes les propriétés étudiées à l'exception de la P_7. La mesure *Dépendance causale* croit en fonction de l'ensemble des données, contrairement à la mesure *Rappel*.

Le deuxième sous-groupe contient les 3 mesures *Jaccard*, *Cosinus* et *Support*. Les deux premières mesures partagent l'ensemble des propriétés à l'exception de la P_{14}. L'allure de ces mesures est illustrée dans la *figure 3.5*.

Le dernier sous-groupe comprend également 3 mesures *Précision*, *Confiance causale* et *Confirmation causale*. Ces dernières partagent toutes les propriétés étudiées sauf P_3, P_{10} et P_{18}. Elles sont donc toutes linéaires, croissantes en fonction du nombre d'exemples et de la taille de l'ensemble des données et décroissantes en fonction du nombre de conséquent. Elles possèdent des valeurs variables dans le cas de l'indépendance, l'équilibre et l'implication logique et sont incapables d'identifier les zones d'attraction et de répulsion.

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	**Classes**
0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	Jaccard
0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	Support
0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	Cosinus
1	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	Rappel
1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	Confirmation causale
1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1	Confiance causale
1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	Dépendance causale
0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1	Précision
0 ?	1	0	1 ?	0	1 ?	0	0	0	0	0	?	0	0	0	0 ?	0	0	1	Classe C_4

TABLE *3.5: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 4.*

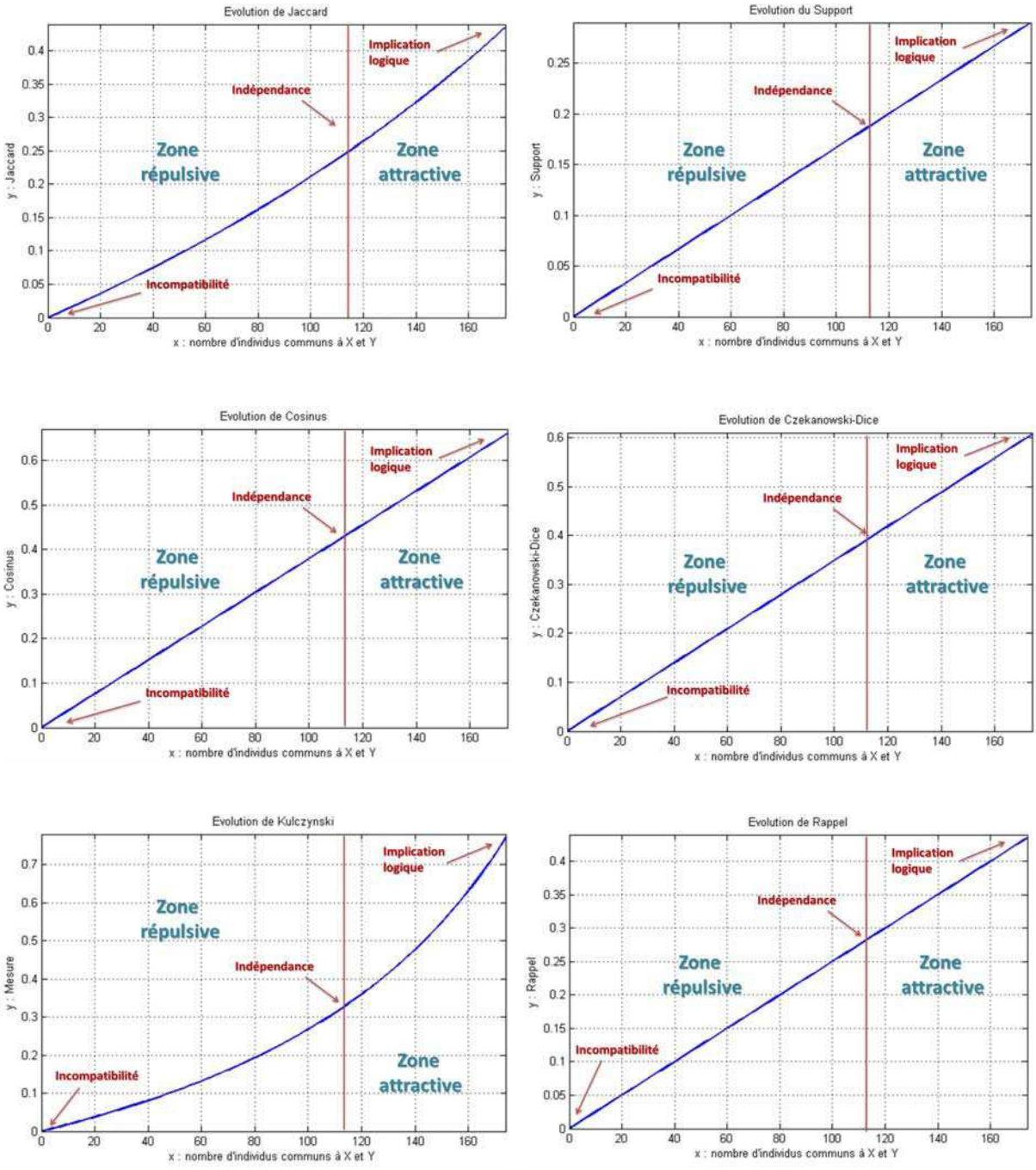

FIGURE *3.5: Évolution de certaines mesures de la classe C_4 en fonction du nombre d'exemples.*

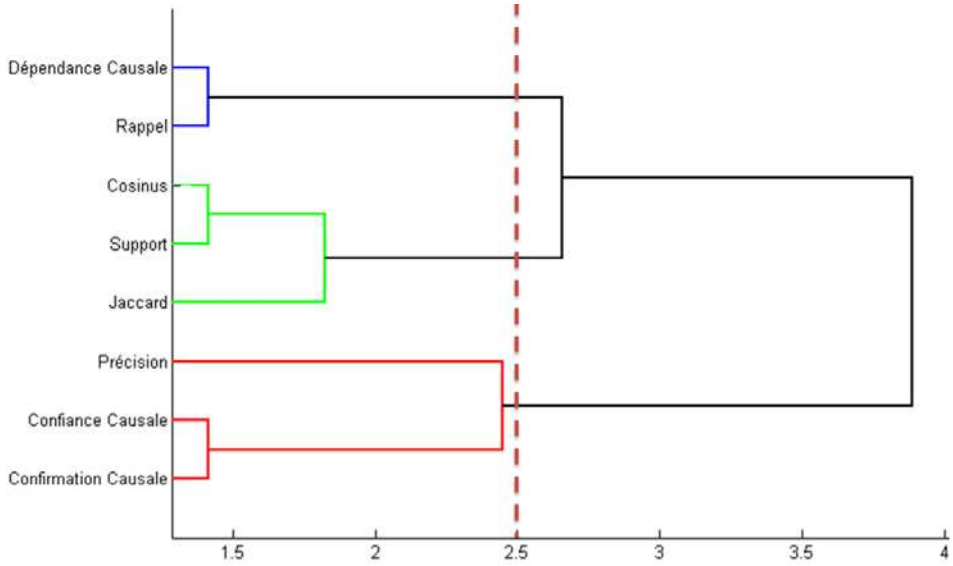

FIGURE *3.6: Classification hiérarchique des mesures de la classe* C_4.

Noyau ou mesure référente de C_4 De même que pour les classes précédentes, nous cherchons entre les différentes mesures de cette classe, celles qui selon la *table 3.10* sont les plus proches du centre de gravité. Nous découvrons que deux mesures peuvent se présenter comme étant les référentes de C_4, à savoir *Rappel* et *Dépendance causale*, appartenant au premier sous-groupe identifié. Ce couple de mesures est distant de $d = 1,32$ par rapport au centre de cette classe. Il est ainsi le sous-groupe le plus proche du centre de gravité. La mesure représentative du deuxième sous-groupe est *Cosinus*, distante du centre pour une valeur égale à $d = 1,72$. Le dernier sous-groupe est représentée par la mesure *Confirmation causale* qui est éloignée du centre de $d = 2,12$.

3.4.4 Étude de la classe C_5

La classe C_5 regroupe 7 mesures d'intérêt {*Confiance, Ganascia, Taux d'exemples, Sebag, Confirmation descriptive, Laplace et Moindre contradiction*}.

– *Étape a* : En consultant les définitions des mesures appartenant à la C_5 (*annexe A*) ainsi que la *table 2.5, page 73*, nous constatons que les indices {*Sebag, Taux d'exemples*} sont liées par une transformation monotone de la *Confiance*. Nous discernons alors les relations suivantes :

- $Sebag = \frac{1}{\frac{1}{Confiance}-1}$
- $Taux\ d'exemples = 2 - \frac{1}{Confiance} = 1 - \frac{1}{Sebag}$

Également, nous remarquons que les mesures *Ganascia et Laplace* peuvent être écrites en fonction de la confiance de la manière suivante :

- $Ganascia = 2 \times Confiance - 1$
- $Laplace = \frac{Confiance \times (T \times P(XY)+1)}{T \times P(XY)+2Confiance}$

Laplace est en fonction de la confiance qui prend en compte le nombre de transactions T. La matrice de distances (*table 3.9*) confirme que ces deux mesures {*Ganascia, Laplace*} sont très proches de la *Confiance* puisque *d(Gan, Conf) = d(Lap, Conf) =* $1,41$.
Selon la *table 2.5, page 73*, d'autres relations existent entre les différentes mesures de cette classe, comme celles entre les mesures *Moindre contradiction et Taux d'exemples*, et entre *Confirmation descriptive et Ganascia* :

- $Moindre\ contradiction = \frac{P(XY)}{P(Y) \times Taux\ d'exemples}$
  La *Moindre contradiction* exprime la différence entre exemples et contre-exemples, et elle est normalisée selon la taille de la conclusion de la règle [Vai06].
- $Confirmation\ descriptive = P(X)\Big(Ganascia - 2\Big)$.

Outre ces relations, la matrice de distance nous révèle la proximité entre la mesure *Moindre contradiction* et les mesures *Sebag, Confirmation descriptive et Laplace*, égale à $d = 1,41$.

– *Étape b* : En regardant la *table 3.6*, nous remarquons que les mesures de cette classe C_5 possèdent 14 propriétés communes (P_3, P_4, P_6, $\overline{P_7}$, $\overline{P_8}$, $\overline{P_9}$, P_{11}, $\overline{P_{12}}$, $\overline{P_{13}}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{18}}$, $\overline{P_{19}}$, $\overline{P_{20}}$). Afin de mieux comprendre les caractéristiques de ces mesures, nous traçons dans la *figure 3.7*, les courbes d'évolution en fonction du nombre d'exemples.
La *figure 3.7* réunit les mesures {*Taux d'exemples, Ganascia, Laplace, Confiance, Sebag, Moindre contradiction et Confirmation descriptive*}. Nous remarquons que les 4 autres mesures se croisent toutes dans un même point au niveau de l'implication logique. Les deux mesures *Laplace* et *Confiance* sont toutes les deux superposées. La mesure *Sebag* est la seule qui possède une allure convexe. Les deux dernières mesures (*Moindre contradiction et Confirmation descriptive*) sont linéaires ($P_{14} = 1$) et se croisent à l'indépendance.
– *Étape c* : Cette étape n'est pas nécessaire pour la classe C_5.

Noyau ou mesure référente de C_5 Nous procédons toujours de la même façon que pour les autres classes afin de découvrir le noyau. Nous visualisons la *table 3.10* et cherchons la mesure la plus proche du centre parmi celles de C_5. Nous trouvons que la mesure *Laplace* est la mesure référente de la classe C_5, pour une distance $d = 1,06$ la séparant du centre.

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	**Classes**
1	1	0	1	0	0	0	1	**1**	0	0	1	0	0	0	0	0	0	1	Confiance
1	1	1	1	0	0	0	**1**	1	0	0	1	0	0	1	0	0	0	1	Ganascia
1	1	1	1	0	0	0	1	1	0	0	**2**	0	0	0	0	0	0	0	Taux d'exemples
1	1	0	**1**	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	Sebag
1	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	Confirmation descriptive
1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	Laplace
1	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	Moindre contradiction
1	1	?	1	0	0	0	?	1	0	0	?	0	0	?	0	0	0	1 ?	Classe C_5

TABLE *3.6: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 5.*

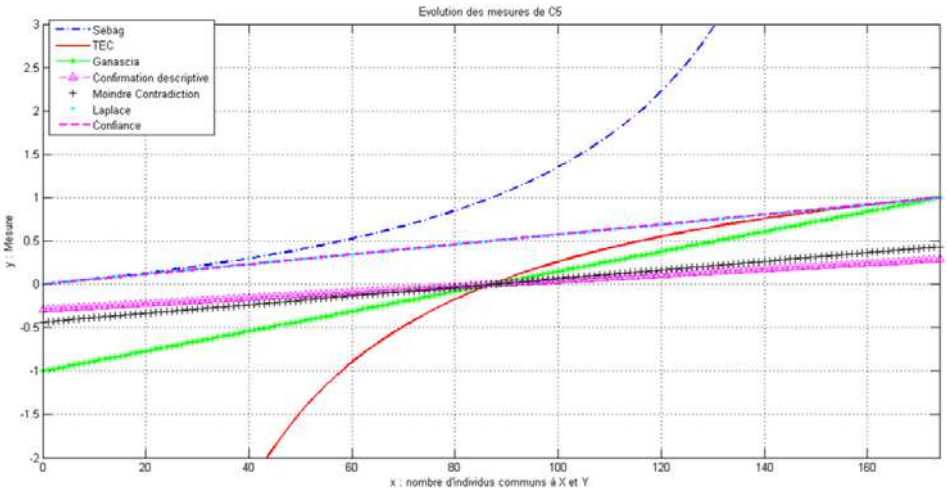

FIGURE *3.7: Évolution des mesures de C_5 en fonction du nombre d'exemples.*

3.4.5 Étude de la classe C_6

La classe C_6 est composée des cinq mesures suivantes : *Zhang*, M_{GK}, *Y et Q de Yule* et *Goodman*.

– *Étape a* : En consultant les définitions des mesures, nous remarquons que les deux mesures *Zhang* et M_{GK} possèdent en numérateur la même mesure *Nouveauté*. Cette dernière appartient à classe C_6 selon la méthode hiérarchique uniquement. En outre, en regardant les définitions des mesures *Y* et *Q de Yule*, rappelées ci-dessous, nous constatons ces deux mesures se ressemblent : l'une (*Y de Yule*) qui prend les paramètres de l'autre (*Q de Yule*) et les met en racine carrée, comme il est montré dans les formules des deux mesures rappelées ci-dessous.

- $Q\ de\ Yule = \frac{P(XY)P(\bar{X}\bar{Y}) - P(X\bar{Y})P(\bar{X}Y)}{P(XY)P(\bar{X}\bar{Y}) + P(X\bar{Y})P(\bar{X}Y)}$
- $Y\ de\ Yule = \frac{\sqrt{P(XY)P(\bar{X}\bar{Y})} - \sqrt{P(X\bar{Y})P(\bar{X}Y)}}{\sqrt{P(XY)P(\bar{X}\bar{Y})} + \sqrt{P(X\bar{Y})P(\bar{X}Y)}}$

– *Étape b* : Selon la *table 3.1*, qui rappelle les caractéristiques de cette classe, toutes les mesures de C_6 vérifient les propriétés P_4, P_5, P_6, P_7, P_9, P_{10}, P_{12}, P_{13}, P_{17} et P_{21}. Grâce

à cet ensemble de propriétés vérifiées, nous pouvons donner une première sémantique à cette classe C_6. Ces mesures sont des indices normalisés puisqu'ils prennent des valeurs fixes pour les cas de l'indépendance ($P_9 = 1$) et de l'implication logique ($P_{10} = 1$) et que les valeurs prises par ces indices permettent de savoir si la règle est dans la zone attractive ($P_{12} = 1$) ou dans la zone répulsive ($P_{13} = 1$).

La *figure 3.8* permet de vérifier cette première sémantique donnée à ces indices. Nous avons tracé l'évolution des cinq mesures lorsque le nombre d'exemples augmente passant ainsi de l'état d'incompatibilité jusqu'à l'implication logique. Nous avons également indiqué sur la *figure 3.8* les trois états caractéristiques d'une règle : l'incompatibilité, l'indépendance et l'implication logique ainsi que les zones d'attraction et de répulsion.

Cette *figure 3.8* ainsi qu'une étude complémentaire nous permettent d'affiner la sémantique apportée à cette classe C_6. Parmi les 61 mesures étudiées, celles de la classe C_6 sont les seules mesures normalisées ayant à la fois des valeurs comprises entre -1 et 1 et des valeurs fixes égales à -1, 0 et 1 pour respectivement l'incompatibilité, l'indépendance et l'implication logique. De plus, elles ont non seulement des valeurs identifiables dans la zone d'attraction et de répulsion mais ces valeurs sont comprises entre 0 et 1 dans la zone d'attraction et entre -1 et 0 dans la zone de répulsion. Pour finir, le signe de la mesure nous renseigne sur la zone d'appartenance de la règle. Nous pouvons donc en déduire que ces mesures évaluent une certaine distance par rapport à l'indépendance : distance entre l'indépendance et l'implication logique dans le cas de valeurs positives et une distance entre l'indépendance et l'incompatibilité dans le cas de valeurs négatives.

Lorsque nous observons la *figure 3.1* révélant la classification ascendante hiérarchique en utilisant le critère de *Ward* ou encore la matrice de distance de la *table 3.9*, nous remarquons une proximité entre les indices *Zhang* et M_{GK} ($d = 2,00$), également entre les indices *Y et Q de Yule* pour une distance *d(YYule, QYule =* $1,41$*)* et entre ce couple de mesures et Goodman ($d = 2,00$). En outre, les symboles du type "?" évoqués dans la *table 3.1* pour les propriétés étudiées, peuvent nous renseigner sur ces deux proximités plus prononcées entre les mesures. La *table 3.7* détaille les différentes propriétés vérifiées par les cinq mesures de ce groupe et rappelle les caractéristiques générales de cette classe. La première propriété où ce symbole apparaît et qui permet d'expliquer ces deux proximités est la symétrie des mesures. Nous avons *Y, Q de Yule* et *Goodman* qui sont des mesures symétriques alors que les mesures *Zhang* et M_{GK} sont des mesures non symétriques.

Les propriétés P_{14} et P_{16} permettent également d'expliquer ces deux proximités. Les indices *Y, Q de Yule* et *Goodman* ont des valeurs opposées pour les règles $X \to Y$ et $\overline{X} \to Y$ et des valeurs identiques pour les règles $X \to Y$ et $\overline{X} \to \overline{Y}$. Les mesures *Zhang*

et M_{GK} vérifient la négation de ces deux dernières propriétés.

– *Étape c* : Cette étape n'est pas nécessaire.

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Classes
1	1	1	1	1	0	1	1	0	1	1	2	1	0	1	0	0	0	1	Zhang
1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	0	0	1	M_{GK}
0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	1	Y de Yule
0	1	1	1	1	0	1	1	0	1	1	2	1	1	1	1	0	0	1	Q de Yule
0	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	0	0	1	Goodman
?	1	1	1	1	0	1	1	0	1	1	?	?	?	1	?	0	0	1	Classe 6

TABLE *3.7: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 6.*

Noyau ou mesure référente de C_6 Nous reprenons la *table 3.10* à la recherche des distances entre les mesures de C_6 et le centre de gravité. La distance minimale égale à $d = 1,84$, est accordée à la mesure *Q de Yule*, qui représente alors la classe C_6.

Nous allons maintenant étudier la classe C_7.

3.4.6 Étude de la classe C_7

La classe 7 contient 10 mesures d'intérêt (*Loevinger ou Facteur de certitude, Pavillon, Facteur bayésien, Conviction, Risque relatif, Gain informationnel, Intérêt, Support double sens, Klosgen et Support sens unique*), dont nous trouvons difficile leur interprétation.

– *Étape a* : Selon la *table 2.5, page 73*, nous avons pu identifié les relations mathématiques suivantes entre certaines mesures de cette classe :

- $Gain\ Informationnel = log_2(Intérêt)$
- $Support\ double\ sens = P(X) \times Support\ sens\ unique$
- $Pavillon = P(\overline{Y}) \times Facteur\ de\ certitude$
- $Klosgen = \sqrt{P(X)} \times pavillon$
- $Facteur\ bayésien = Conviction \times Intérêt$
- $Facteur\ de\ certitude = \frac{\left(Risque\ relatif \times p(Y|\bar{X})\right) - p(Y)}{p(\bar{Y})}$
- $Support\ double\ unique = P(XY) log_2(Intérêt)$

Étant données ces relations, nous remarquons que certains couples de mesures, tel que le couple {*Gain informationnel, Intérêt*}, sont fortement liés et faciles à interpréter. Contrairement à d'autres couples de mesures de cette même classe C_7, qaui ne révèlent aucune relation mathématique intéressante, tel est le cas du couple {*support sens unique, Klosgen*}.

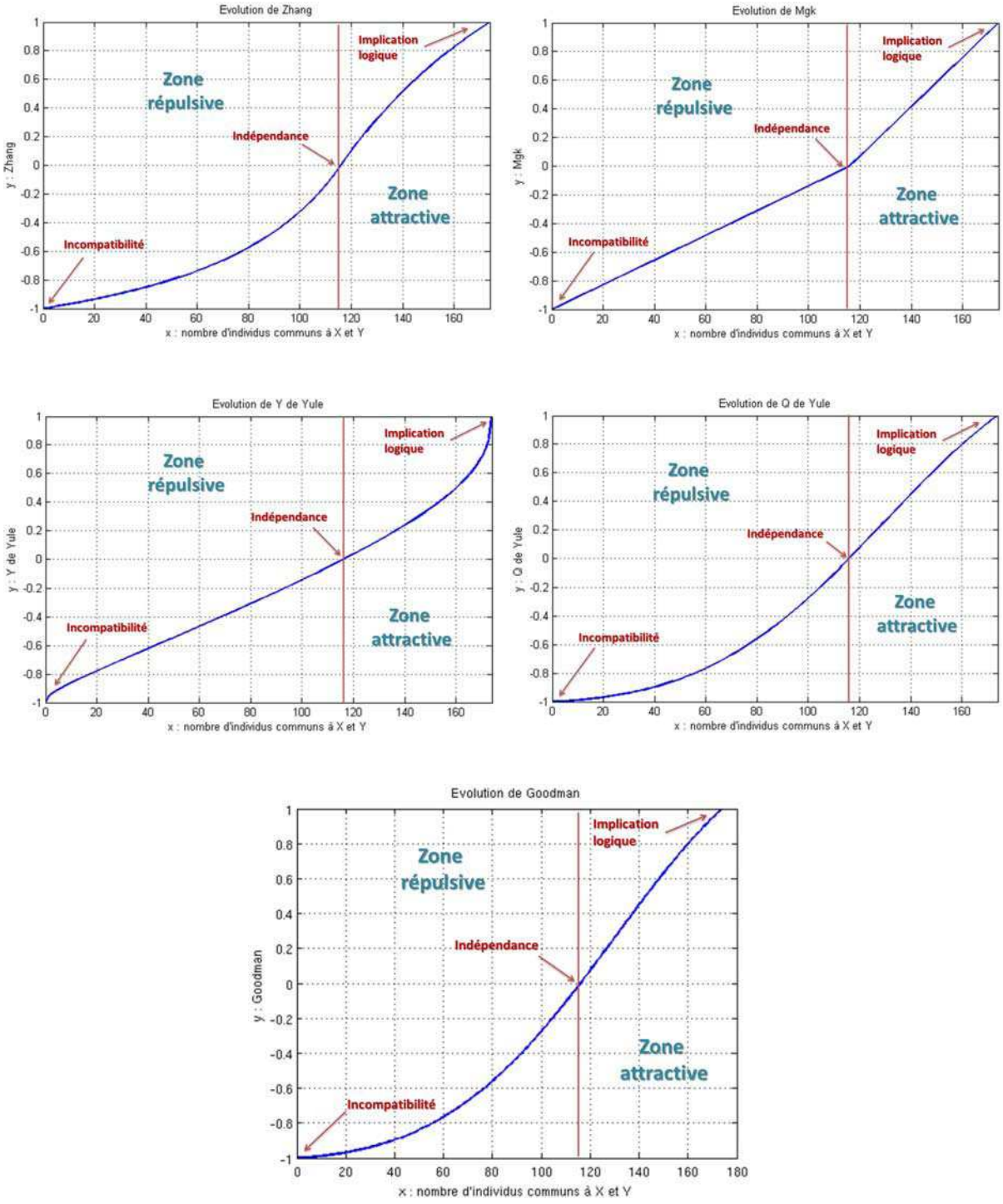

FIGURE *3.8: Évolution des cinq mesures de la classe C_6 en fonction de la variation du nombre d'exemples.*

– *Étape b* : D'après la *table 3.8*, toutes les mesures de la classe C_7 vérifient les propriétés suivantes : P_4, P_7, P_9, $\overline{P_{11}}$, P_{12}, P_{13}, $\overline{P_{16}}$, $\overline{P_{18}}$, $\overline{P_{19}}$, $\overline{P_{20}}$ et P_{21}. Elles sont donc des mesures descriptives, discriminantes, qui croient en fonction de l'ensemble des données, qui ne possèdent pas de valeur fixe à l'équilibre et qui permettent d'identifier les zones d'attraction et de répulsion.
– *Étape c* : Face à la difficulté d'interprétation de cette classe et au nombre relativement élevé de ses mesures, nous appliquons encore une fois la classification hiérarchique utilisant le critère de Ward, sur le tableau disjonctif complet correspondant aux mesures de C_7. Le dendrogramme de la *figure 3.9* nous révèle les quatre sous-groupes de mesures suivants :

  (1) G_{c1} qui regroupe les mesures *Facteur de certitude* et *Pavillon*,

  (2) G_{c2} qui comprend *Conviction*, *Facteur bayésien* et *Risque relatif*,

  (3) G_{c3} qui contient les mesures *Intérêt* et *Gain informationnel*,

  (4) G_{c4} qui regroupe les mesures restantes *Klosgen*, *Support sens unique* et *Support double sens*.

  Le premier sous-groupe identifié G_{c1} contient deux mesures révélant une dépendance mathématique (*voir étape a*) où il s'agit de passer de *Facteur de certitude* à *Pavillon* en multipliant par une valeur $P(\overline{Y})$ non nulle. D'après la *table 3.8*, ces deux mesures évaluent 18 propriétés de la même manière à l'exception de la P_5, P_{10} et P_{17}.
  Le deuxième sous-groupe G_{c2} comprend deux mesures *Facteur bayésien* et *Conviction*, qui selon la *table 3.8* sont très proches puisqu'elles évaluent 19 propriétés de la même manière. Elles sont aussi proches de la mesure *Intérêt* selon la formule présente lors de l'*étape a*. En termes de distances entre ces mesures (*table 3.9*), nous trouvons que *d(FB, Conv) =* $2,00$ *et d(FB ou Conv, Int) =* $2,45$. La différence entre ces deux valeurs peut justifier la non présence de la mesure *Intérêt* au sein de G_{c2} mais confirme sa proximité à ce couple de mesures. Selon la *table 3.8*, seule la mesure *Intérêt* est linéaire ($P_{14} = 1$) contrairement aux deux autres qui sont convexes ($P_{14} = 0$) privilégiant ainsi les contre-exemples. La *figure 3.10 (gauche)* présente les courbes correspondantes à ces mesures et étudie leur évolution en fonction du nombre d'exemples. Quant à la mesure *Risque relatif* appartenant aussi à ce sous-groupe, elle est proche des mesures *Facteur bayésien* et *Conviction* avec 17 propriétés communes parmi 19 (*table 3.8*). La *figure 3.10 (gauche)* vient prouver la ressemblance du comportement de ces trois mesures puisqu'elles sont parallèles de l'incompatibilité jusqu'à l'indépendance.
  Au sein du troisième sous-groupe G_{c3}, nous retrouvons les mesures *Gain informationnel* et *Intérêt* qui sont liées par une relation logarithmique. Selon la *table 3.10*, ces deux

P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	Classes
1	1	0	1	1	1	1	0	0	1	1	0	1	0	0	0	0	0	1	Facteur bayésien
1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	1	Pavillon
1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	Conviction
1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	1	Loevinger
1	1	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	Risque relatif
0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1	Intérêt
0	1	0	1	1	1	1	0	0	1	1	**2**	0	0	0	0	0	0	1	Gain informationnel
1	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1	Klosgen
1	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	Support sens unique
0	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	Support double sens
?	1	?	?	1	1 ?	1	0 ?	0	1	1	?	0 ?	0	0 ?	0	0	0	1	Classe 7

TABLE *3.8: Évaluation des propriétés d'un sous-ensemble de mesures de la classe 7.*

mesures possèdent 19 propriétés communes parmi 21. La *figure 3.10* vient également confirmer le comportement similaire de ce couple de mesures puisque toutes les deux croient en fonction du nombre d'exemples et sont parallèles à partir de l'indépendance. En effet, la mesure *Gain informationnel* possède une allure concave ($P_{14} = 2$), tandis que la mesure *Intérêt* est linéaire ($P_{14} = 1$). La matrice de distance de la *table 3.9* prouve aussi la proximité des deux mesures avec une distance *d(Int, GI)* = $1,41$.
Le dernier sous-groupe G_{c4} de la classe C_7 contient trois mesures "*Support sens unique*, *Support double sens* et *Klosgen*". Nous avons pu identifier une relation mathématique entre les deux premières mesures uniquement (*voir étape a*). Le comportement similaire de ce couple de mesures est aussi prouvé par la *table 3.8*, qui montre que toutes les deux évaluent toutes les propriétés de la même manière à l'exception de la P_3. Également, la mesure *Support sens unique* est proche de *Klosgen* avec *18* propriétés communes parmi *19*. Cette concordance au niveau de l'évaluation des propriétés par les trois mesures confirme leur regroupement. Quant à la matrice de distance (*table 3.9*), elle réaffirme ce rapprochement en déterminant des valeurs de distance entre ces trois mesures égales à *d(SSU, SDS) = d(SSU, Klos)* = $1,41$.

Noyau ou mesure référente de C_7 Toujours nous procédons de la même façon que pour toutes les autres classes. Nous consultons la *table 3.10* pour identifier les distances entre les mesures de C_7 et le centre de gravité de cette classe. La mesure qui possède la distance minimale est *Risque relatif* avec $d = 0,78$. Elle est donc la mesure représentative de la classe C_7. Cette mesure appartient à G_{c2} qui est donc le sous-groupe le plus proche du centre, suivi du sous-groupe G_{c4}, dont la mesure représentative est *Support sens unique* distante du centre pour $d = 1,58$. Par la suite, nous retrouvons les sous-groupes G_{c3} puis G_{c1}. Chacun de ces sous-groupes peut être représenté par l'une des mesures leur appartenant.

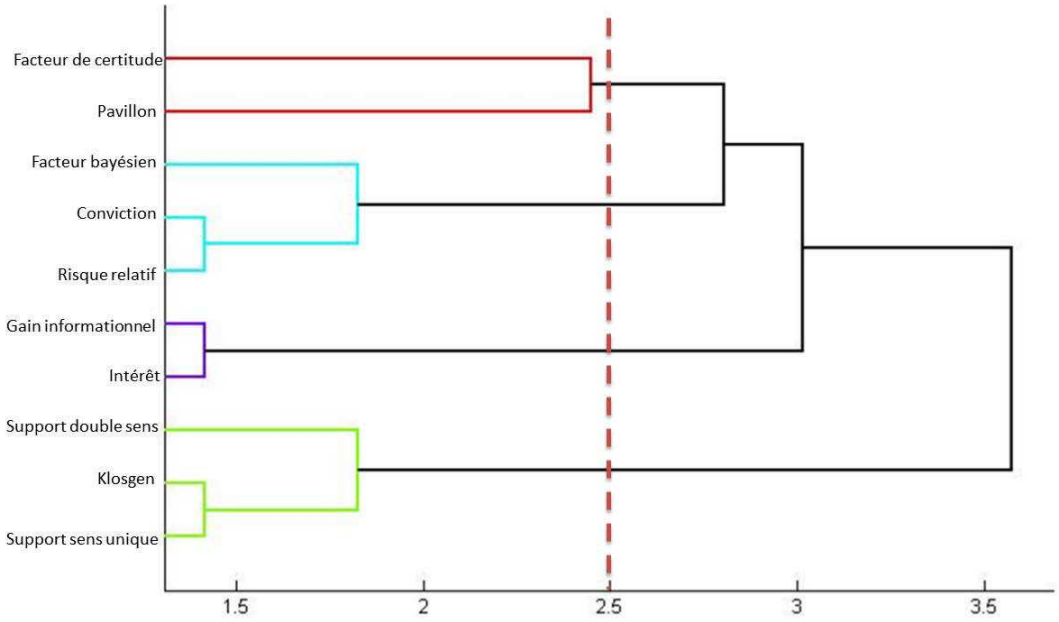

FIGURE *3.9: Classification hiérarchique des mesures de la classe* C_7.

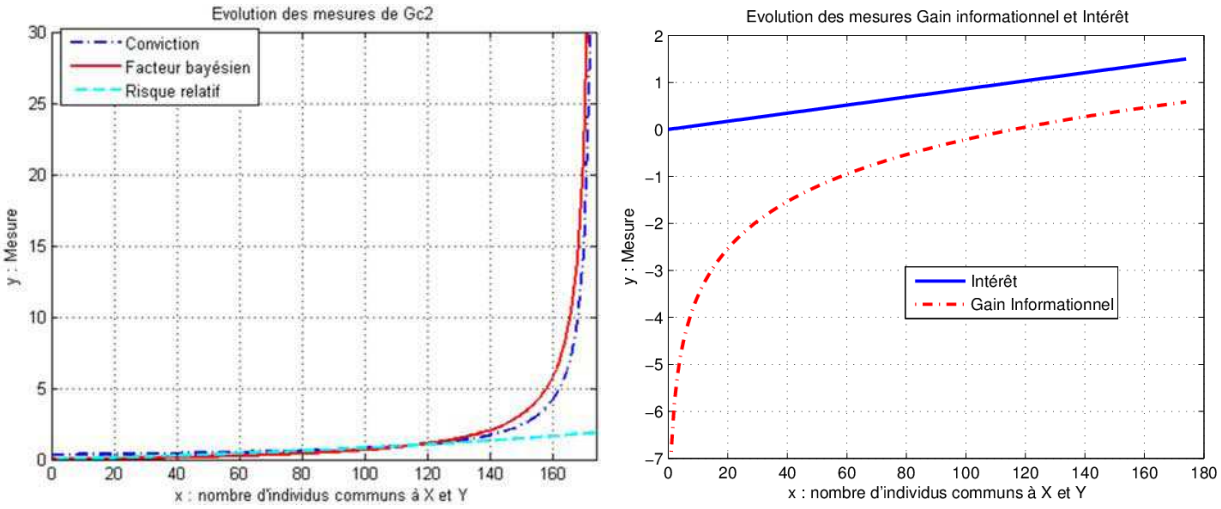

FIGURE *3.10: Évolution des mesures de* G_{c2} *(gauche) et de* G_{c3} *(droite) en fonction du nombre d'exemples de la classe* C_7.

3.4.7 Étude des mesures instables

Dans ce qui suit, nous nous intéressons à l'étude des mesures instables, *i.e.,* celles dont aucun consensus n'a été révélé par les méthodes de classification et qui n'étaient pas aussi présentes dans les classes fortes de la *section 3.3.4*. Nous discernons les *11* mesures instables suivantes : {*VT100, Piatetsky-Shapiro, Coefficient de corrélation, Force collective, Cohen, Dépendance pondérée, Fukuda, Prévalence, Précision, J-mesure et Gini*}. Comme nous l'avons déjà signalé, les cinq premières mesures forment le groupe G_{p8}, identifié dans la *sous-section*

3.3.3 par une version courante de l'algorithme des *k-moyennes*. Cependant, les résultats révélés par la *CAH* catégorisent ces mesures différemment. Nous discutons au fur et à mesure ces regroupements.

Nous commençons par la mesure *VT100* qui appartient à la classe C_4 selon la méthode hiérarchique. La matrice de distance de la *table 3.9* montre que cette dernière possède des valeurs de distance assez élevées avec les autres mesures. Nous trouvons que les deux mesures qui lui sont les plus proches sont *Précision* de la classe C_4 avec une distance *d(VT100, Préc)* = $2,45$ et le *Coefficient de corrélation* du groupe G_{p8} avec une distance égale à $2,83$. Cela peut expliquer sa déviation vers ces deux groupes.

De même, nous constatons aussi que la mesure *Précision* n'est pas toujours stable et qu'elle peut dévier avec les mesures du groupe G_{p8} pour les mêmes raisons, sa distance avec *VT100* qui est pratiquement égale à sa distance avec les autres mesures de C_4. Néanmoins, il est possible de justifier d'avantage la déviation des mesures *Précision et VT100* de leurs groupes par une simple visualisation de la *table 3.10*. Cette dernière indique que ces deux mesures sont les plus éloignées du centre de leurs classes. Nous trouvons que la mesure *Précision* est distante du centre de C_4 pour une valeur élevée de $d = 4,12$ et de même pour *VT100*, qui est éloignée du centre de G_{p8} pour $d = 5,92$. Cela justifie que ces deux mesures se trouvent en périphérie de leurs classes et par conséquent, elles peuvent dévier facilement vers d'autres classes aussi proches d'elles, en appliquant les méthodes de classification.

Les mesures *Piatetsky-Shapiro (PS) et Coefficient de corrélation (Cor)* sont toujours ensemble quelque soit le résultat de la classification (*même après plusieurs exécutions de l'algorithme des k-moyennes*). Cela est expliqué par leur distance *d(Cor, PS)* = $1,41$, qui montre qu'elles sont très proches. Ces deux mesures se regroupent parfois avec les mesures de la classe C_6 à cause de leur rapprochement avec la mesure *Goodman* pour *d(Cor, Good)* = $2,00$. Elles sont autant proches des mesures du groupe G_{p8} qu'à la mesure *Goodman*.

Également, les mesures *Force collective (FCol) et Cohen (Coh)* sont toujours unies quelque soit la méthode classification appliquée, toutes les deux sont séparées par une distance *d(FCol, Coh)*= $1,41$. La matrice de distance indique que ces deux mesures sont d'une part, proches des mesures *Intérêt, Gain Informationnel et Conviction* de C_7, et d'autre part du *Coefficient de corrélation* du groupe G_{p8}, pour une même distance $d \leq 2,45$.

La *table 3.10* montre que la mesure *Coefficient de corrélation* peut représenter le groupe G_{p8} puisqu'elle est la plus proche de son centre de gravité.

Nous passons maintenant à l'étude des mesures *Dépendance pondérée (DP) et Fukuda (Fuk)*. Ces dernières appartiennent à la classe C_3 selon la classification hiérarchique et à la classe C_4 selon la méthode des *k-moyennes*. L'instabilité de ce couple de mesures peut-être expliquée par les valeurs élevées des distances qui les séparent des mesures des classes

C_3 et C_4, sachant que toutes les deux sont distantes pour une valeur de *d(DP, Fuk)* = $2,83$. Néanmoins, nous trouvons que *Dépendance pondérée* est proche des mesures *Dépendance causale et Cosinus* pour $d = 2,00$ et que *Fukuda* est proche de la mesure *Support* pour $d = 2,00$.

Les mesures restantes {*Prévalence, J-mesure*} appartiennent à la classe C_3. Cette classe ne fait pas partie des classes fortes puisque ses mesures ne sont pas toujours groupées ensemble après les multiples exécutions de l'algorithme des *k-moyennes*. Parfois, nous trouvons la *J-mesure* groupée avec le couple {*Variation support, Pearl*} pour une distance qui vaut respectivement $2,45$ et $2,83$. Ces valeurs s'expliquent par la visualisation de la ligne correspondante à la *J-mesure* dans la matrice de distance (*table 3.9*), où nous remarquons des valeurs assez élevées avec la plupart des mesures à l'exception des mesures *Gini, Dépendance et Couverture* de la classe C_3. Dans d'autres cas, nous découvrons une déviation de la mesure *Prévalence* avec la classe C_4. Cela est expliqué par sa distance avec les mesures *Rappel* et *Support* de C_4 ($d = 2,45$). Cependant, elle reste toujours plus proche de la mesure *Couverture* appartenant aussi à la classe C_3 ($d = 1,41$). Finalement, la mesure *Gini* est regroupée le plus souvent avec la mesure qui lui est la plus proche, la *J-mesure*.

Après avoir étudié les 7 classes, les mesures instables et tenté de donner une interprétation à celles-ci, nous validons maintenant notre travail par une comparaison avec des classifications existant dans la littérature, celles dégagées par Vaillant [Vai06], Huynh et al. [HGB+07], Heravi et Zaïane [HZ10], Le Bras [Bra11], Lesot et Rifqi [LR10], et pour finir Zighed et al. [ZAB11].

3.5 Étude comparative avec les autres travaux : Validation

Cette section s'intéresse à la dernière étape de notre processus d'analyse, qui est la comparaison des résultats de la classification que nous avons obtenue dans la *sous-section 3.3* avec les classifications de [Vai06], [HGB+07], [LR10], [Bra11], [ZAB11] et [HZ10]. Nous commençons par une comparaison avec les résultats dégagés par Vaillant [Vai06].

3.5.1 Comparaison avec le travail de Vaillant

Vaillant [Vai06] a mené son étude sur 20 mesures, dont 19 mesures communes, selon 9 propriétés formelles. Sur ces 9 propriétés, nous en avons 7 en commun puisque nous avons écarté les propriétés "*compréhensibilité de la mesure*" et "*facilité à fixer un seuil d'acceptation*", jugées trop subjectives. Pour effectuer sa classification, Vaillant a également appliqué la méthode hiérarchique (*CAH*) en utilisant le critère de Ward mais il a retenu la distance de Manhattan. L'auteur fait remarquer qu'en utilisant la distance euclidienne, il a obtenu des

résultats semblables. Il a dégagé les 5 classes suivantes :

- $ClBV_1$ = *{Support, Moindre contradiction, Laplace}*,
- $ClBV_2$ = *{Confiance, Sebag, Taux d'exemples}*,
- $ClBV_3$ = *{Coefficient de corrélation, Piatetsky-Shapiro, Pavillion, Intérêt, Indice d'implication, Cohen, Gain informationnel}*,
- $ClBV_4$ = *{Loevinger ou Facteur de certitude, Facteur bayésien, Conviction}* et
- $ClBV_5$ = *{Zhang, Intensité d'implication, Indice probabiliste discriminant (IPD)}*.

En confrontant nos deux résultats, nous dégageons un accord sur les regroupements suivants :

$ClBV_2 \subset C_5$, $ClBV_4 \subset C_7$,

et nous avons les relations suivantes entre les groupes obtenus par Vaillant et ceux que nous avons identifiés précédemment :

$ClBV_1 - Support \subset C_5$, $ClBV_3 - Indice\ d'implication \subset Gp_8 \cup C_7$ et $ClBV_5 - Zhang \subset C_1 \cup C_2$.

Le groupement où le désaccord est le plus important est le groupe $ClBV_3$ puisque nous avons dû faire apparaître le groupe Gp_8 présent uniquement avec l'une des techniques : celle des *k-moyennes*. Quant au groupe $ClBV_5$, il regroupe, à l'exception de la mesure *Zhang*, toutes les mesures de la famille de l'intensité d'implication. Cette divergence des résultats peut être expliquée par les 12 propriétés supplémentaires que nous avons étudiées.

Une autre classification est réalisée par [HGB+07] sur les mesures d'intérêt sera présentée dans la sous-section suivante et comparée avec les 7 groupes de mesures que nous avons obtenus.

3.5.2 Comparaison avec le travail de Hyunh et al.

Une autre classification a été effectuée par Huynh et al. [HGB+07], qui ont étudié 36 mesures d'intérêt, dont 32 mesures communes. Les auteurs présentent dans un premier temps une taxonomie des mesures selon les 2 critères suivants :

1. *le sujet* : la déviation de l'indépendance ou de l'équilibre ;
2. *la nature* : descriptive ou statistique.

À partir de l'étude des mesures sur ces 2 critères, les 5 groupes de mesures suivants sont obtenus, où nous ne retenons que les mesures communes :

- Cl_{de} (*descriptive/déviation de l'équilibre*) : {*Confiance, Laplace, Sebag, Taux d'exemples, Confirmation descriptive, Confiance confirmée-descriptive, Moindre contradiction*} ;

- Cl_{di} (*descriptive/déviation de l'indépendance*) : {*Corrélation, Intérêt, Loevinger, Conviction, Dépendance, Pavillon, J-mesure, Gini, Force collective, Ratio des chances, Q de Yule, Y de Yule, Klosgen, Cohen*} ;
- Cl_{se} (*statistique/déviation de l'équilibre*) : {*IPEE}* ;
- Cl_{si} (*statistique/déviation de l'indépendance*) : {*II, IIE*} ;
- Cl_o (*autres*) : {*Support, Précision, Jaccard, Cosinus, Confiance causale, Confirmation causale, Confiance confirmée-causale, Dépendance causale*}.

En confrontant ces 5 groupes de mesures à ceux décrits dans la *figure 3.2, page 89*, nous remarquons notre accord sur la catégorisation des mesures suivantes :

{*Confiance, Laplace, Sebag, Taux d'exemples, Moindre contradiction*} $\subset C_5$,

{*Corrélation, Cohen, Force collective, Ratio des chances*} qui sont groupées ensemble selon la méthode de partitionnement des*k-moyennes*,

{*Gini, J-mesure, dépendance, klosgen*} $\subset C_5$,

{*Intérêt, Loevinger, Conviction, Pavillon, Klosgen}* $\subset C_7$,

{*Q de Yule, Y de Yule*} $\subset C_6$,

{*Jaccard, Cosinus, Confirmation causale, Confiance causale, Confiance confirmée-causale*} $\subset C_4$.

Étant donnée cette confrontation, nous remarquons une grande similarité au niveau de la classification des mesures communes.

Dans ce qui suit les résultats des travaux de classification réalisés par Heravi et Zaiane [HZ10] sont passés en revue et comparés avec les 7 classes de mesures que nous avons obtenues.

3.5.3 Comparaison avec les travaux de Heravi et Zaiane

L'étude effectuée par Heravi et Zaiane [HZ10] a comporté 53 mesures d'intérêt objectives dont 45 mesures sont communes. Cette étude a permis l'obtention d'un dendrogramme pour catégoriser les mesures suite à l'application de la méthode hiérarchique *CAH* sur l'ensemble de ses données (53 *mesures* $\times$ 16 *propriétés*).

Afin de comparer nos travaux, nous coupons le dendrogramme illustré dans [HZ10] à un niveau où le nombre des classes est proche de celui que nous avons obtenu, qui est le nombre 7. Nous gardons seulement les mesures communes et nous retenons les 12 groupes de mesures suivantes :

- Cl_{HZ1} : {*Support sens unique, J-mesure, Support double sens, Dépendance pondérée, Klosgen*} ;
- Cl_{HZ2} : {*Gini, Information mutuelle, Goodman*} ;

- Cl_{HZ3} : {*Intensité d'implication*} ;
- Cl_{HZ4} : {*Variation support double sens, Corrélation, Piatetsky-shapiro*} ;
- Cl_{HZ5} : {*Ratio des chances, Q de Yule, Y de Yule*} ;
- Cl_{HZ6} : {*Précision, Cohen, Gain informationnel, Intérêt*} ;
- Cl_{HZ7} : {*Pavillon, Risque relatif, Confirmation causale, Leverage, Facteur de certitude, Spécificité, Force collective*} ;
- Cl_{HZ8} : {*Cosinus, Czekanowski-dice, Jaccard, Rappel*} ;
- Cl_{HZ9} : {*Zhang, Facteur bayésien, Indice d'implication*} ;
- Cl_{HZ10} : {*Confiance causale, Confiance confirmée-causale, Conviction, Loevinger*} ;
- Cl_{HZ11} : {*Confiance, Ganascia, Confiance confirmée-descriptive, Sebag, Taux d'exemples, Moindre contradiction*} ;
- Cl_{HZ12} : {*Confirmation descriptive, Laplace, Support*}.

Tout d'abord, nous signalons quelques remarques que nous avons pu identifier sur ces 12 groupes de mesures : nous trouvons que les mesures *Loevinger* et *Facteur de certitude* appartiennent respectivement aux classes Cl_{HZ10} et Cl_{HZ7}. Cependant, nous avons vu précédemment qu'il existe des mesures ayant une même définition mais avec des noms différents, d'où ce couple de mesures {*Loevinger, Facteur de certitude*}. Les auteurs ne donnent pas dans leur article les expressions des mesures, ce qui ne nous permet pas de voir si leurs définitions sont correctes. Nous retrouvons aussi le couple de mesures {*Confirmation descriptive, Ganascia*} au sein du groupe Cl_{HZ11}, qui de même possèdent des noms différents pour une même définition.

Néanmoins, par la confrontation de nos deux travaux, nous constatons une cohérence au niveau de ces regroupements de mesures :

Le premier groupe Cl_{HZ1} constitue deux ensembles de mesures appartenant à deux classes différentes. Nous avons {*Support sens unique, Support double sens, Klosgen*} $\subset C_7$ et {*J-mesure, Dépendance pondérée*} $\subset C_3$ selon la méthode de classification hiérarchique ;

Le second groupe définit la relation suivante : $Cl_{HZ2}-$ {*Goodman*} $\subset C_3$;

Le groupe Cl_{HZ3} ne contient qu'une seule mesure, *Intensité d'implication* ;

Le groupe $Cl_{HZ4}-$ {*Variation support double sens*} $\subset G_{p8}$, groupe identifié par la méthode des *k-moyennes* ;

Le groupe $Cl_{HZ5}-$ {*Ratio des chances*} $\subset C_6$;

Le groupe $Cl_{HZ6}-$ {*Précision*} $\subset C_7$. La mesure *Cohen* appartient à C_7 selon la méthode de classification hiérarchique ;

Les mesures du groupe Cl_{HZ7} sont issues des deux classes C_4 et C_7. Nous avons {*Pavillon, Risque relatif, Facteur de certitude, Force collective*} $\subset C_7$ et les mesures {*Confirmation causale, Leverage, Spécificité*} $\subset C_4$;

Le groupe $Cl_{HZ8} \subset C_4$;

Les mesures du groupe Cl_{HZ9}, {*Zhang, Facteur bayésien, Indice d'implication*}, appartiennent respectivement aux classes C_6, C_7 et C_3 ;

Les mesures {*Confiance causale, Confiance confirmée-causale*} de Cl_{HZ10} appartiennent à la classe C_4, tandis que les mesures {*Conviction, Loevinger*} appartiennent à C_7.

Le groupe $Cl_{HZ11} \subset C_5$;

Finalement, nous avons le groupe $Cl_{HZ12}-$ *Support* $\subset C_5$.

À partir de cette comparaison, nous constatons que les classes C_4 et C_7 ont été divisées en des sous-classes et que la classe C_5 est pratiquement vérifiée si nous coupons le dendrogramme à un niveau plus haut. Nous expliquons les différences existant entre nos classifications par la présence de mesures additionnelles qui ont été étudiées par les auteurs.

Dans ce qui suit, nous allons comparer nos résultats avec ceux obtenus par Y. Le Bras [Bra11].

3.5.4 Comparaison avec le travail de Le Bras

Le Bras [Bra11], [BLL12] cherche dans ses travaux à trouver des caractéristiques communes aux mesures objectives. Pour ce faire, il a étudié 42 mesures d'intérêt selon 6 critères opérationnels qu'il a proposés. Ces critères concernent d'une part la possibilité de calculer la robustesse [BMLL10b], et d'autre part d'utiliser des algorithmes efficaces. Les 6 critères étudiés sont : $Br_{1.1}$, $Br_{1.2}$, $Br_{2.1}$, $Br_{2.2}$, $Br_{2.3}$, Br_3, rappelés dans le *chapitre 1* (*cf. sous-section 1.6.2.6, page 32*).

Au vu de ces 42 mesures, nous notons au total 38 mesures communes, dont certaines d'entre elles possèdent une même définition mais avec des noms différents [4]. Par la confrontation de nos travaux, nous cherchons à identifier si les mesures communes, qui sont groupées ensemble selon notre étude formelle, possèdent aussi un même comportement selon [Bra11].

La comparaison de chaque groupe de mesures C_1 à C_7 de la *figure 3.2* avec les résultats obtenus par l'étude de Le Bras révèle les similarités suivantes :

– C_3 : les mesures communes aux deux travaux qui appartiennent à ce groupe sont {*Couverture, Gini, Indice d'implication, J-mesure, Prévalence et Pearl*[5]}. Selon l'auteur, ces mesures ne sont ni quadratiques ni anti-monotones. Il montre également la proximité des mesures *Couverture* et *Prévalence*, puisqu'elles sont les seules mesures planes et omni-monotones vérifiant la propriété GUEUC ;

4. Intérêt représente la mesure *Pearl* avec nos travaux, *Levier* représente la mesure *Nouveauté* et *J1-mesure* correspond à la mesure *Support double sens*

5. *qui appartient à C_3 selon la méthode des k-moyennes*

- C_4 : contient les mesures communes {*Cosinus, Czekanowski-dice, Jaccard, Kulczynski, Précision, Spécificité, Support* et *Rappel*}. Toutes ces mesures, à l'exception de la mesure qui est quadratique, sont planes et vérifient la propriété d'antimonotonie. En outre, nous remarquons que la majorité de ces mesures possèdent la propriété GUEUC, à l'exception de *Kulczynski* et *Spécificité*. La mesure *Support* est la seule à être omni-monotone dans ce groupe ;
- C_5 : nous remarquons que seule la mesure*Confirmation descriptive* est absente de ce groupe. L'étude réalisée par Le Bras [Bra11] découvre que les mesures {*Taux d'exemples, Sebag, Ganascia* et *Confiance*} vérifient toutes les propriétés analysées de la même manière. Toutefois, aucune des mesures communes de C_5 n'est quadratique. Cependant, elles sont toutes omni et opti-monotones. Nous réalisons aussi que toutes ces mesures sont planes, à l'exception de *Laplace* et que seulement deux de ces mesures (*Moindre contradiction et Laplace*) sont anti-monotones ;
- C_6 : nous identifions trois mesures opti-monotones *Y, Q de Yule* et *Zhang*, qui ne vérifient aucune de ces propriétés, antimonotonie, omni-monotonie et mesure plane. Cependant, en visualisant le comportement des mesures *Piatetsky-shapiro* et *Nouveauté*, qui appartiennent à cette classe selon la méthode hiérarchique, nous trouvons qu'ils sont aussi opti-monotones et ne vérifient pas la propriété d'omni-monotonie et de mesure plane. La *Nouveauté*, qui semble être plus robuste que la mesure *Piatetsky-Shapiro* (*qui est quadratique*), est la seule mesure qui possède la meilleure propriété d'anti-monotonicité dans le cas des règles de classe ;
- C_7 : toutes les mesures de C_7 ont été étudiées par l'auteur, incluant les mesures {*Force collective, Cohen et Ratio des chances*}, qui selon la méthode hiérarchique appartiennent à C_7. Parmi toutes ces mesures, seule *Cohen* est anti-monotone, mais aucune n'est omni-monotone ou plane. La propriété GUEUC est vérifiée par les mesures {*Pavilion, Conviction, Facteur bayésien, Gain informationnel, Intérêt*} et *Loevinger*, qui sont aussi quadratiques et opti-monotones, identifiant par conséquent les propriétés opérationnelles fortes avec les mesures *Cohen*, *Ratio des chances* et *Risque relatif*.

Suite à cette comparaison, nous remarquons qu'à partir de l'étude de Le Bras sur les mesures d'intérêt selon les six critères proposés, nous trouvons que la plupart des mesures communes aux deux travaux vérifient les critères étudiés par Le Bras, comme c'est expliqué précédemment.

Dans ce qui suit, nous confrontons nos résultats avec ceux de [LR10], [ZAB11].

3.5.5 Comparaison avec les autres travaux

D'autres classifications ont été réalisées sur des mesures de distance et de similarité par Lesot et Rifqi [LR10] ainsi que par [ZAB11].

3.5.5.1 Comparaison avec le travail de Lesot et Rifqi

Dans [LR10], les auteurs ont étudié l'ordre induit par les mesures et non pas les valeurs numériques obtenues puisque leur contexte d'étude est la recherche d'information. Cette étude a porté sur des mesures dédiées aux données binaires et aux données numériques en réalisant des expérimentations sur à la fois des données réelles et sur des données artificielles. Les auteurs ont obtenu une liste de mesures équivalentes (*mesures qui induisent toujours le même ordre*) et pour les mesures non équivalentes, les auteurs ont quantifié ce désaccord par un degré d'équivalence basé sur le coefficient de Kendall généralisé. Sur les 10 mesures étudiées et destinées aux données binaires, 5 sont communes à nos deux études. Ces mesures sont les suivantes : *Czekanowski-Dice, Jaccard, Ochiai* et *Q et Y de Yule*. Les auteurs ont trouvé que les mesures *Q et Y de Yule* sont des mesures équivalentes. Ce résultat est également confirmé par notre étude puisque ces deux mesures sont dans la classe C_6 et comme nous l'avons déjà mentionné, elles sont très proches sur le dendrogramme de la *figure 3.1*. Les auteurs ont également trouvé que les mesures *Czekanowski-Dice* et *Jaccard* sont des mesures équivalentes. Ces deux mesures ont été également affectées à la même classe : la classe C_4, et nous les retrouvons avec une proximité relativement importante dans le dendrogramme de la *figure 3.1* (*nous avons choisi la mesure Cosinus comme mesure représentative sur le dendrogramme puisque comme nous l'avons vu dans la section 3.3.1, les mesures Cosinus et Czekanowski-Dice ont des valeurs identiques pour les* 19 *propriétés ce qui a conduit à la constitution du groupe* G_3). Pour finir, nous avons regroupé également la mesure *Ochiai* (*ou Cosinus*) avec les mesures *Czekanowski-Dice* et *Jaccard*, dans la classe C_4. Les auteurs [LR10] ont trouvé un degré d'équivalence entre la mesure *Cosinus* et la classe d'équivalence {*Czekanowski-Dice, Jaccard*} de $0,99$, ce qui conforte nos résultats.

3.5.5.2 Comparaison avec le travail de Zighed et al

Une dernière classification a été proposée par Zighed et al. [ZAB11] sur 13 mesures de proximité dont seulement deux mesures sont communes à nos deux études : le *Cosinus* et le *Coefficient de corrélation*. Cette classification proposée par les auteurs est basée sur l'équivalence topologique et fait appel à la structure de voisinage local. Les deux mesures sont apparues très proches dans cette classification, contrairement à nos travaux puisque nous les retrouvons dans les classes C_4 et G_{p8}. L'ensemble des mesures étudiées étant

tellement différent, les classes trouvées par chacune des techniques ne peuvent donc être que difficilement comparables. En outre, comme ça été souligné par les auteurs, la classification qu'ils ont obtenue est faiblement représentative puisqu'elle fût réalisée sur un seul jeu de données : les *Iris de Fisher*.

Nous sommes bien conscients que la catégorisation des mesures peut aussi dépendre de plusieurs facteurs parmi lesquels : les données, l'expert-utilisateur, la nature des règles extraites et la procédure de recherche des classes, comme le souligne Suzuki [Suz08].

Afin d'éviter le biais des données, de l'expert et de la nature des règles extraites, nous avons ici fait le choix d'une étude théorique basée sur des propriétés de mesures [GGM10], plutôt que sur des données expérimentales [HGB05a]. Les deux aspects sont bien évidemment complémentaires. Pour éviter le biais de la procédure de construction de classes, nous avons utilisé deux techniques de classification qui, de manière générale, ont exhibé de fortes ressemblances entre de nombreuses mesures, et fait ressortir des similitudes et des différences avec des travaux précédents [Vai06], [LR10], [ZAB11].

Cette étude vient compléter des travaux précédents sur la description d'une vision unificatrice des mesures d'intérêt [HC07], et apporte une contribution supplémentaire à l'analyse de ces mesures.

3.6 Conclusion

Ce chapitre a pris comme point de départ le travail de synthèse sur les mesures d'intérêt réalisé dans le *chapitre 2*. Ce travail de synthèse a conduit à l'obtention d'une matrice d'évaluation de 19 propriétés jugées intéressantes sur 61 mesures. L'objectif de ce chapitre est la classification de ces mesures afin d'élucider les ressemblances entre les mesures d'intérêt, et contribuer à aider l'utilisateur dans le choix de mesures complémentaires au couple {*Support, Confiance*}, capables d'éliminer les règles inintéressantes. Dans un premier temps, nous avons analysé ces données (*matrice de* 61 *mesures* $\times$ 19 *propriétés*) afin de déterminer si une simplification n'était pas envisageable. Ainsi, nous avons recherché tout d'abord tous les groupes de mesures au comportement totalement identique, ensuite nous avons détecté si des propriétés n'étaient pas redondantes. Nous avons identifié 7 groupes de mesures au comportement totalement identique ce qui a permis de réduire nos données de départ. Une matrice beaucoup moins importante que celle d'origine a été proposée permettant d'aider d'avantage l'utilisateur dans son choix de mesures. Par exemple, dans le cas où l'utilisateur souhaite des mesures très différentes ou des mesures ayant des propriétés particulières, son choix sera facilité avec la consultation de cette table.

Ayant la matrice de données réduite après l'élimination de la redondance, nous avons effectué une classification grâce à deux techniques : une méthode de la classification ascendante hiérarchique et une version de la méthode des *k-moyennes*. Les classifications obtenues grâce aux deux techniques ont permis de trouver d'une part des classes fortes, identifiées suite à de multiples exécutions de l'algorithme des *k-moyennes* et d'autre part un consensus, où 7 classes ont été identifiées. Ces classes ont été en partie validées par des travaux de classification existant dans la littérature. Dans la *table 3.11*, nous positionnons notre travail par rapport aux études formelles existant.

Ainsi, l'étude réalisée dans ce chapitre sur la classification des mesures d'intérêt a permis l'obtention de groupes de mesures disjoints. Face à cette caractéristique, nous procédons dans le chapitre suivant à appliquer une des méthodes de classification avec recouvrement pour obtenir une nouvelle structure de la classification avec des groupes de mesures qui se chevauchent.

Points clés

Positionnement :

- *Classification d'une soixantaine de mesures d'intérêt en utilisant des méthodes sans recouvrement*.

Contribution :

- *Classification des mesures en utilisant deux méthodes de classification non supervisée : CAH et k-moyennes* ;
- *Obtention de* 7 *classes de mesures au comportement identique* ;

Publications :

- *S. Guillaume, D. Grissa, E. Mephu Nguifo (2011). Catégorisation des mesures d'intérêt pour l'extraction des connaissances*. Dans EGC'2011, pages 551–562.
- *S. Guillaume, D. Grissa, E. Mephu Nguifo (2011). Catégorisation des mesures d'intérêt pour l'extraction des connaissances*. Dans Revue des Nouvelles Technologies de l'Information, RNTI. pages 117–144.

		1	2	3	4	5	6	8	9	10	11	12	14	15	16	17	18	20	21	22	23	24	25	26	27	28	30	31	32	33	34	35	36	37	39	41	42	44	45	46	47	48	49	51	52	54	55	56	57	58	59	60	61
1	Cor	0,00																																																			
2	Coh	2,00	0,00																																																			
3	Conf	4,69	4,24	0,00																																																	
4	CConf	4,00	3,46	2,45	0,00																																																
5	Pav	2,83	2,83	4,24	3,46	0,00																																																
6	Gan	4,47	4,47	1,41	2,83	4,00	0,00																																														
8	CfmC	3,74	3,16	2,83	1,41	3,16	3,16	0,00																																													
9	CfmD	4,47	4,47	2,45	3,46	3,46	2,00	3,16	0,00																																												
10	Conv	3,16	2,45	4,00	3,16	2,45	4,24	2,83	4,24	0,00																																											
11	Cos	4,00	3,46	3,16	2,83	3,46	3,46	2,45	2,83	3,74	0,00																																										
12	Couv	5,10	4,69	3,46	3,74	4,24	3,74	3,46	3,16	4,47	3,16	0,00																																									
14	Dep	4,69	4,24	3,46	3,74	3,74	3,74	3,46	3,16	4,00	3,16	2,00	0,00																																								
15	PCD	4,00	3,46	3,16	2,00	2,83	3,46	1,41	2,83	3,16	2,00	3,16	3,16	0,00																																							
16	DP	4,69	4,24	3,46	3,16	3,74	3,74	2,83	3,16	3,74	3,16	2,83	2,83	2,45	0,00																																						
17	FB	3,74	3,16	4,47	3,74	2,45	4,69	3,46	4,24	2,00	3,74	4,47	4,00	3,16	3,74	0,00																																						
18	FC	3,16	2,45	3,46	2,45	2,45	3,74	2,83	4,24	2,00	3,74	4,47	4,00	3,16	4,00	2,83	0,00																																						
20	FCol	2,45	1,41	4,47	3,74	3,16	4,69	3,46	4,69	2,00	3,74	4,90	4,47	3,74	4,24	2,83	2,83	0,00																																						
21	Fuk	4,69	4,24	2,83	3,16	3,74	3,16	2,83	2,45	4,00	2,45	2,83	2,83	2,45	2,83	4,00	4,00	4,47	0,00																																						
22	GI	3,16	2,45	4,47	3,74	2,45	4,69	3,46	4,24	2,45	3,16	4,47	4,00	3,16	3,46	2,45	2,83	2,45	4,00	0,00																																						
23	Gini	4,90	4,47	3,74	4,00	4,47	4,00	3,74	3,46	4,24	3,46	2,45	2,45	3,46	2,83	4,24	4,69	4,24	3,16	4,47	0,00																																						
24	Good	2,00	2,83	4,24	4,00	3,46	4,00	4,24	4,47	3,74	4,47	5,10	4,69	4,47	4,69	4,24	3,16	3,16	4,69	3,74	4,90	0,00																																						
25	IndImp	4,47	4,00	3,16	3,46	4,00	3,46	3,16	3,46	3,74	3,46	3,16	2,45	3,46	3,16	4,24	3,74	4,24	2,45	4,24	3,46	4,47	0,00																																						
26	IPEE	5,29	4,90	3,16	4,00	4,90	3,46	3,74	3,46	4,47	4,00	3,74	4,24	4,00	3,74	4,90	4,69	4,90	3,16	4,69	4,24	5,29	3,46	0,00																																						
27	IP3E	5,10	4,69	2,83	3,74	4,69	3,16	3,46	3,16	4,24	3,74	4,00	4,00	3,74	3,46	4,69	4,47	4,69	2,83	4,47	4,00	5,10	3,16	1,41	0,00																																						
28	IPD	4,69	4,24	4,00	3,74	4,69	4,24	3,46	4,24	4,24	3,16	4,47	4,47	3,74	4,00	4,24	4,47	4,24	3,46	4,00	4,47	5,10	3,74	3,16	2,83	0,00																																						
30	II	4,24	3,74	4,90	4,24	3,74	5,10	4,00	5,10	3,16	4,69	4,90	4,90	4,24	4,47	3,16	3,46	3,74	4,47	3,46	5,29	4,69	4,24	3,74	4,00	3,46	0,00																																						
31	IIE	4,69	4,24	3,46	3,16	4,24	3,74	2,83	3,74	3,74	3,74	4,00	4,00	3,16	2,83	4,24	4,00	4,24	2,83	4,00	4,00	4,69	3,16	2,45	2,00	2,83	3,46	0,00																																						
32	IIER	4,00	3,46	4,24	4,00	3,46	4,47	3,74	4,47	2,83	4,47	4,69	4,24	4,00	3,74	3,46	3,16	3,46	3,74	3,16	4,69	4,00	3,46	3,46	3,16	3,74	2,45	2,45	0,00																																						
33	IVL	3,74	3,16	4,90	4,24	3,74	5,10	4,00	5,10	3,16	4,24	4,90	4,90	4,24	4,47	3,74	3,46	3,16	4,47	2,83	5,29	4,24	4,24	3,74	4,00	3,46	2,00	3,46	2,45	0,00																																						
34	Int	2,83	2,00	4,24	3,46	2,00	4,47	3,16	4,00	2,45	2,83	4,24	3,74	2,83	3,74	2,45	2,45	2,45	3,74	1,41	4,47	3,46	4,00	4,90	4,69	4,24	3,74	4,24	3,46	3,16	0,00																																						
35	Jac	4,24	3,74	3,46	3,16	3,74	3,74	2,83	3,16	3,46	1,41	3,46	3,46	2,45	3,16	3,46	4,00	3,46	2,83	3,16	3,16	4,69	3,74	4,00	3,74	3,16	4,69	3,74	4,47	4,24	3,16	0,00																																						
36	Jmes	4,69	4,24	4,00	4,24	3,74	4,24	4,00	3,74	3,46	3,74	2,83	2,00	3,74	3,16	3,46	4,00	4,00	3,46	3,74	2,45	4,69	3,16	4,47	4,24	4,69	4,69	4,24	4,00	4,69	3,74	3,46	0,00																																						
37	Klos	4,00	3,46	4,24	4,00	2,83	4,47	3,74	4,00	2,45	4,00	3,74	3,16	3,46	2,83	2,45	3,16	3,16	3,74	2,83	3,46	4,00	3,46	4,69	4,47	4,90	4,00	4,00	3,16	4,00	2,83	3,74	2,45	0,00																																						
39	Lap	4,47	4,00	1,41	2,83	4,00	2,00	2,45	2,00	3,74	2,83	3,16	3,16	2,83	3,16	4,24	3,74	4,24	2,45	4,24	3,46	4,47	2,83	2,83	2,45	3,74	4,69	3,16	4,00	4,69	4,00	3,16	3,74	4,00	0,00																																						
41	Mgk	3,16	3,16	3,46	3,16	2,45	3,16	3,46	3,74	2,83	4,24	4,47	4,00	3,74	4,00	3,46	2,00	3,46	4,00	3,46	4,69	2,45	3,74	4,69	4,47	4,90	4,00	4,00	3,16	4,00	3,16	4,47	4,00	3,16	3,74	0,00																																						
42	MC	4,69	4,24	2,00	3,16	3,74	2,45	2,83	1,41	4,00	2,45	2,83	2,83	2,45	2,83	4,00	4,00	4,47	2,00	4,00	3,16	4,69	3,16	3,16	2,83	4,00	4,90	3,46	4,24	4,90	3,74	2,83	3,46	3,74	1,41	4,00	0,00																																						
44	Pearl	3,74	3,16	4,00	4,24	4,24	4,24	4,00	4,24	4,00	3,74	3,46	2,83	4,24	4,00	4,47	4,00	3,46	4,00	4,00	3,16	3,74	3,16	4,69	4,47	4,47	4,90	4,47	4,24	4,47	3,74	4,00	2,83	3,74	3,74	4,00	4,00	0,00																																						
45	PS	1,41	2,45	4,90	4,24	3,16	4,69	4,00	4,69	3,46	4,24	5,29	4,90	4,24	4,90	4,00	3,46	2,83	4,47	3,46	5,10	2,45	4,24	5,10	4,90	4,47	4,00	4,47	3,74	3,46	3,16	4,47	4,90	4,24	4,69	3,46	4,90	4,00	0,00																																						
46	Prec	3,16	2,45	3,46	2,45	3,74	3,74	2,00	3,74	3,46	2,45	4,00	4,00	2,45	3,46	4,00	3,46	2,83	3,46	3,46	3,74	3,74	3,74	4,24	4,00	3,46	4,47	3,46	4,24	4,00	3,16	2,83	4,47	4,24	3,16	4,00	3,46	3,46	3,46	0,00																																						
47	Prev	4,90	4,47	3,16	3,46	4,00	3,46	3,16	2,83	4,24	2,83	1,41	2,45	2,83	2,45	4,24	4,24	4,69	2,45	4,24	2,83	4,90	2,83	3,46	3,74	4,24	4,69	3,74	4,47	4,69	4,00	3,16	3,16	3,46	2,83	4,24	2,45	3,74	5,10	3,74	0,00																																						
48	YQ	2,83	3,46	4,69	4,47	4,00	4,47	4,69	4,90	4,00	4,90	5,48	5,10	4,90	4,69	4,00	3,74	3,46	5,10	3,74	5,10	2,00	4,90	5,29	5,10	4,69	4,24	4,69	4,00	4,24	4,00	4,90	4,90	4,24	4,90	3,16	5,10	4,24	3,16	4,24	5,29	0,00																																						
49	Rap	4,24	3,74	2,83	2,45	3,16	3,16	2,00	2,45	3,46	1,41	2,83	2,83	1,41	2,83	3,46	3,46	4,00	2,00	3,46	3,16	4,69	3,16	3,74	3,46	3,46	4,47	3,46	4,24	4,47	3,16	2,00	3,46	3,74	2,45	4,00	2,00	4,00	4,47	2,83	2,45	5,10	0,00																																						
51	RR	3,46	2,83	4,24	3,46	2,00	4,47	3,16	4,00	1,41	3,46	4,24	3,74	2,83	3,46	1,41	2,45	2,45	3,74	2,00	4,00	4,00	4,00	4,69	4,47	4,47	3,46	4,00	3,16	3,46	2,00	3,16	3,16	2,00	4,00	3,16	3,74	4,24	3,74	3,74	4,00	4,24	3,16	0,00																																						
52	Seb	4,90	4,47	2,45	3,46	4,00	2,83	3,16	2,00	3,74	2,83	3,16	3,16	2,83	2,83	3,74	4,24	4,24	2,45	4,00	2,83	4,90	3,46	3,16	2,83	4,00	4,90	3,46	4,24	4,90	4,00	2,45	3,16	3,46	2,00	4,24	1,41	4,24	5,10	3,74	2,83	5,10	2,45	3,46	0,00																																						
54	Sup	4,24	3,74	2,83	3,16	3,74	3,16	2,83	2,45	4,00	1,41	2,83	2,83	2,45	2,83	4,00	4,00	4,00	2,00	3,46	3,16	4,24	3,16	3,74	3,46	3,46	4,90	3,46	4,24	4,47	3,16	2,00	3,46	3,74	2,45	4,00	2,00	3,46	4,47	2,83	2,45	4,69	2,00	3,74	2,45	0,00																																						
55	SSU	3,74	3,16	4,47	3,74	2,45	4,69	3,46	4,24	2,00	3,74	4,00	3,46	3,16	3,16	2,00	2,83	2,83	4,00	2,45	3,74	4,24	3,74	4,90	4,69	4,69	3,74	4,24	3,46	3,74	2,45	3,46	2,83	1,41	4,24	3,46	4,00	4,00	4,00	4,00	3,74	4,47	3,46	1,41	3,74	4,00	0,00																																						
56	SDS	3,46	2,83	4,69	4,00	2,83	4,90	3,74	4,47	2,45	3,46	4,24	3,74	3,46	3,46	2,45	3,16	2,45	4,24	2,00	4,00	4,00	4,00	5,10	4,90	4,47	4,00	4,47	3,74	3,46	2,00	3,16	3,16	2,00	4,47	3,74	4,24	3,74	3,74	3,74	4,00	4,24	3,74	2,00	4,00	3,74	1,41	0,00																																						
57	TEC	5,10	4,69	2,00	3,16	4,69	2,45	3,46	3,16	4,24	3,74	3,46	4,00	3,74	3,46	4,69	4,00	4,69	3,46	4,47	4,00	4,69	3,74	2,45	2,83	4,00	4,47	3,46	4,24	4,47	4,69	3,74	4,24	4,47	2,45	4,00	2,83	4,47	5,29	4,00	3,16	4,69	3,46	4,47	2,83	3,46	4,69	4,90	0,00																																						
58	VT100	2,83	3,46	4,24	3,46	4,00	4,00	3,16	4,00	4,24	3,46	4,69	4,69	3,46	4,24	4,69	4,24	3,74	4,24	4,24	4,47	3,46	4,47	4,47	4,24	3,74	4,69	3,74	4,47	4,24	4,00	3,74	5,10	4,90	4,00	4,24	4,24	4,24	4,24	3,16	2,45	4,47	4,00	3,74	4,47	4,47	3,74	4,69	4,47	4,69	0,00																																						
59	VS	4,00	3,46	4,24	4,47	4,47	4,47	4,24	4,47	3,74	4,00	3,74	3,16	4,47	4,00	4,24	4,24	3,16	4,24	4,00	2,83	4,00	3,46	4,69	4,47	4,47	4,90	4,47	4,24	4,47	4,00	3,74	2,45	3,46	4,00	4,24	4,24	1,41	4,24	3,74	4,00	4,24	4,24	4,00	4,00	3,74	3,74	3,46	4,47	4,47	0,00																																						
60	YY	2,83	3,46	4,69	4,47	4,00	4,47	4,69	4,90	3,74	4,90	5,48	5,10	4,90	4,90	3,74	3,74	3,16	5,10	4,00	4,90	2,00	4,90	5,48	5,29	4,90	4,47	4,90	4,24	4,47	4,00	4,69	4,69	4,00	4,90	3,16	5,10	4,24	3,16	4,24	5,29	1,41	5,10	4,00	4,90	4,69	4,24	4,00	4,90	4,00	4,00	0,00																																						
61	Zhang	3,74	3,74	4,00	3,74	3,16	3,74	4,00	4,24	3,16	4,69	4,90	4,47	4,24	4,00	3,16	2,83	3,74	4,47	3,46	4,90	3,16	4,24	4,69	4,47	4,47	3,46	4,00	3,16	4,00	3,74	4,69	4,24	3,46	4,24	2,00	4,47	4,47	4,00	4,47	4,69	2,45	4,47	3,46	4,47	4,47	3,74	4,00	4,00	4,69	4,47	2,83	0,00																																						

TABLE *3.9: Matrice de distances entre les mesures. Les abréviations des différentes mesures se trouve dans la table C.1, page 210.*

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	G_{p8}
II	1,56	11,13	19,56	17,12	21,63	15,04	10,58	13,92
IVL	1,56	11,13	18,56	15,52	21,63	15,04	9,78	9,92
IPEE	11,56	2,13	13,56	11,92	7,63	23,44	20,58	21,52
IP3E	12,22	1,13	12,56	9,92	6,20	21,44	18,58	19,52
IPD	10,89	4,13	15,56	9,52	14,20	20,64	17,38	15,52
IIE	8,22	2,13	12,56	7,92	10,20	17,44	14,58	15,52
Couverture	21,56	14,13	3,31	7,72	8,78	23,44	15,78	21,52
Dépendance	20,22	15,13	1,81	7,72	9,35	19,44	11,78	18,32
Gini	24,22	15,13	3,31	8,72	10,49	21,44	15,18	18,72
JMesure	18,22	17,13	2,81	11,12	12,49	17,84	9,18	18,32
IndiceImplication	14,22	9,13	4,81	8,12	9,06	17,44	12,98	15,52
Prévalence	19,56	12,13	4,31	5,72	6,78	21,44	13,78	19,52
ConfianceCausale	15,56	11,13	11,81	3,72	7,35	13,44	10,18	11,52
ConfirmationCausale	13,56	9,13	9,81	2,12	7,06	15,44	8,58	9,52
Cosinus	18,22	11,13	8,31	1,72	7,35	19,04	10,18	11,52
DépendanceCausale	15,56	11,13	9,31	1,32	7,35	17,44	7,38	11,52
Jaccard	18,22	11,13	8,81	3,12	8,49	19,44	9,58	12,72
Précision	16,22	12,13	11,31	4,12	11,06	14,64	11,38	5,52
Rappel	17,56	10,13	7,31	1,32	5,35	19,44	9,38	13,52
Support	18,89	10,13	6,31	2,52	5,35	17,04	11,78	13,52
Confiance	20,22	9,13	9,81	6,52	1,35	15,44	15,78	17,52
Laplace	18,22	7,13	7,81	4,92	1,06	17,44	14,18	15,52
Ganascia	22,22	11,13	11,81	8,52	2,20	13,44	17,38	17,12
ConfirmationDescriptive	22,22	11,13	9,31	6,12	2,20	17,44	14,58	17,12
TauxExemple	17,56	8,13	12,06	9,92	3,92	17,44	18,18	21,12
MoindreContradiction	20,22	9,13	7,31	4,12	1,35	19,44	12,98	17,52
Sebag	20,22	9,13	7,81	5,52	2,49	19,84	12,38	18,72
Zhang	10,89	17,13	17,06	15,92	15,35	3,04	9,38	13,12
MGK	12,22	18,13	13,81	12,52	12,20	3,44	7,38	9,52
QYule	15,56	22,13	20,56	19,92	21,35	1,84	14,18	8,72
YYule	17,56	24,13	19,81	19,92	21,35	2,24	13,18	8,32
Goodman	16,89	23,13	17,31	16,52	18,20	2,24	12,18	5,12
SupSensUnique	11,56	19,13	9,81	10,32	16,49	13,84	1,58	11,12
SupDoubelSens	12,22	20,13	10,81	10,72	18,49	13,44	2,38	9,12
Conviction	7,56	15,13	12,31	9,52	14,20	9,84	1,98	7,12
Pavillon	11,56	19,13	13,31	9,32	14,20	9,44	2,98	7,52
FacteurBayésien	10,22	18,13	13,81	10,72	16,49	11,44	2,38	11,12
FacteurCertitude	9,56	17,13	13,81	9,72	13,35	7,44	4,18	7,92
GainInformationnel	8,22	16,13	13,56	9,12	16,49	11,04	2,58	7,52
Intérêt	10,22	18,13	12,31	7,72	15,35	11,04	2,18	5,92
Klosgen	12,22	18,13	7,81	11,12	14,49	11,84	3,18	13,12
RisqueRelatif	9,56	17,13	11,81	8,72	14,49	11,84	0,78	9,12
VT100	18,22	14,13	17,31	10,12	15,92	14,24	16,98	5,92
Correlation	14,22	22,13	17,31	14,12	19,92	6,24	8,98	1,12
Cohen	10,22	18,13	13,31	10,12	17,06	8,64	5,38	1,92
ForceCollective	10,22	18,13	13,81	11,52	18,20	9,04	4,78	3,12
PiatetskyShapiro	12,22	20,13	18,81	15,72	21,92	8,24	10,98	2,32
VariationSuppport	18,89	18,13	4,81	13,92	16,20	15,04	13,18	12,32
Pearl	18,89	18,13	4,31	12,52	15,06	14,64	13,78	11,12
DépendancePondérée	16,22	10,13	6,56	5,12	8,49	17,44	10,18	17,12
Fukuda	16,22	7,13	6,81	3,72	5,35	19,44	12,98	16,72
IIER	2,22	8,13	14,56	13,92	16,20	11,44	8,58	11,92

TABLE *3.10: Matrice de distances entre les mesures et les centres des classes.*

Auteurs	Nbre mesures	Nbre prop	Nature des propriétés	Nature de l'étude	Techniques utilisées	Résultats
Piatetsky-shapiro [PS91a]	–	3	Propriétés sur la nature des mesures	Proposition de propriétés	–	3 nouvelles propriétés de mesures
Tan et al. [TKS02]	21	8	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Proposition de propriétés + évaluation des mesures selon les propriétés	–	Étude comparative des mesures selon les propriétés
Lallich et Teytaud [LT04]	15	13	Propriétés sur la nature des mesures	Proposition de propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Étude comportementale des mesures
Blanchard et al. [BGBG5a] [BGBG5c]	19	4	Propriétés sur la nature des mesures	Proposition de propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Une nouvelle propriété de mesures ainsi qu'une nouvelle mesure sont proposées et 4 catégories de mesures ont été identifiées
Geng et Hamilton [GH07]	38	11	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Proposition de propriétés + évaluation des mesures selon les propriétés	–	Étude synthétique sur les mesures d'intérêt
Maddouri et Gammoudi [MG07]	62	12	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés	–	Étude théorique sur les mesures d'intérêt
B. Vaillant [Vai06]	20	9	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés + catégorisation des mesures	Classification ascendante hiérarchique	Identification de 5 groupes de mesures
X.-H. Hyunh [Huy06]	36	5	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Identification de 5 groupes de mesures
D. Feno [Fen07]	15	13	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Étude des propriétés des mesures + catégorisation des mesures	Catégorisation selon M_{GK} normalisable	Identification de 3 catégories de mesures
Heravi et Zaiane [HZ10]	53	11	Propriétés portant à la fois sur la nature des mesures et sur leur formulation des matrices	Catégorisation des mesures	Classification ascendante hiérarchique	Identification de groupes de mesures
Y. Le Bras [Bra11], [BLL12]	42	6	Propriétés algorithmiques + propriétés sur la nature des mesures	Proposition de propriétés + évaluation des mesures selon les propriétés + catégorisation des mesures	Catégorisation selon des critères bien déterminés	Identification de 6 catégories de mesures
Notre contribution	61	22	Propriétés sur la nature des mesures	Évaluation des mesures selon les propriétés + Catégorisation des mesures	Classification ascendante hiérarchique + méthode des *k-moyennes*	Identification de 7 classes de mesures après un consensus sur la classification

TABLE *3.11: Tableau de synthèse positionnant notre contribution par rapport à l'ensemble des travaux réalisés sur les mesures d'intérêt selon une approche formelle.*

CHAPITRE 4

Classification des mesures d'intérêt : méthode avec recouvrement

Sommaire

4.1 Introduction

Pour catégoriser les mesures d'intérêt, nous avons mené une première expérimentation dans le *chapitre 3*, où nous avons proposé une structure disjointe des mesures. Néanmoins, les méthodes que nous avons appliquées ne sont pas adaptées à la recherche d'une organisation des données en classes recouvrantes. Or, ce type de schéma de classification devient pourtant indispensable pour conserver une synthèse riche d'informations. Par exemple, certaines mesures peuvent appartenir à plusieurs classes, tel que le couple de mesures {*Fukuda, Dépendance pondérée*}, dont aucun consensus n'a été trouvé par la classification non recouvrante et qui peut appartenir soit à C_3, soit à C_4. Ainsi, et afin de découvrir des recouvrements possibles à partir de nos données d'entrée, nous menons une seconde expérimentation qui sera conduite sur l'ensemble des mesures d'intérêt et de leurs propriétés. Chaque mesure sera étiquetée par une ou plusieurs étiquettes parmi un ensemble de catégories identifiées par l'application de méthode de classification non supervisée avec recouvrement.

Ces catégories de mesures permettront d'une part, de fournir une vue structurée sur les nombreuses mesures d'intérêt existantes. Une telle représentation nous rend plus conscients des similitudes et, éventuellement, des relations existant entre les mesures. D'autre part, d'aider l'utilisateur à choisir la (les) mesure(s) qui répond(ent) à ses attentes. Chaque classe révélée peut en effet, représenter un genre particulier de mesures intéressantes. De ce point de vue, chaque classe identifiée représente ainsi une collection de mesures ayant des propriétés communes, d'où différentes propositions de règles d'association intéressantes sont suggérées. Si un utilisateur a par exemple l'intention de tenir compte de toutes les suggestions émises par ces classes, il peut opter pour la sélection d'une mesure par cluster. C'est le cas dans un article récent de Bouker et al. [BSYN12], où les auteurs proposent de générer des règles d'association par le biais de différentes mesures d'intérêt par agrégation des valeurs des mesures choisies.

Tous les travaux de la littérature [Vai06], [Huy06], [Fen07], [HZ10], [Bra11] réalisés jusqu'à présent sur la classification des mesures d'intérêt, y compris les résultats que nous avons obtenus dans le *chapitre 3*, ont tous une caractéristique commune : utilisation de méthodes sans recouvrement, qui ne produisent que des groupes disjoints de mesures.

Toutefois, étant données que les classes sont censées correspondre à des catégories importantes de mesures, on peut naturellement s'attendre à des groupes qui se chevauchent plutôt qu'à des groupes disjoints. À savoir, une mesure donnée peut appartenir à la fois à deux classes. Elle peut partager certaines propriétés avec des mesures de l'une des classes et, en même temps, partager d'autres propriétés avec des mesures de l'autre classe. Le but de ce chapitre est d'explorer la possibilité d'obtenir des groupes raisonnables de mesures qui se chevauchent.

Dans ce chapitre, nous commençons par définir la classification avec recouvrement, et rappeler ses méthodes ainsi que décrire la méthodologie suivie. Nous introduisons par la suite la méthode de classification avec recouvrement que nous allons appliquer. Ensuite, nous présentons notre étude de cas sur l'utilisation des facteurs booléens pour la classification des mesures, nous interprétons les classes obtenues et nous les comparons avec les résultats de la classification disjointe. Enfin, nous confrontons les deux résultats de classifications à la recherche des groupes de mesures stables.

4.2 Classification avec recouvrement

Cette section définit la classification avec recouvrement, présente ses différentes méthodes à partir desquelles nous allons sélectionner celle(s) la/les mieux appropriée(s) à nos données, et enfin met en oeuvre la méthodologie de travail que nous allons suivre tout au long de ce chapitre.

4.2.1 Méthodes de classification

Dans le cadre de ce chapitre, nous nous intéressons au problème de classification des mesures d'intérêt en utilisant une méthode avec recouvrement. Ce type de méthodes, à l'inverse des méthodes de classification non supervisée classiques (*étudiées dans le chapitre 3*), permet d'obtenir des classes recouvrantes.

Méthodes avec recouvrement Les recherches menées autour de la classification avec recouvrements des classes sont motivées par des besoins apparaissant dans des différents domaines d'application (*les documents multimédia, les données biologiques, etc.*), où des données peuvent appartenir à une ou plusieurs catégories. La classification avec recouvrement est une technique de clustering qui permet à un élément (*ou objet ou individu*) d'être membre de plusieurs groupes à la fois. En la comparant aux techniques de classification non supervisée qui répartissent les données en régions non recouvrantes [HYN+12], nous trouvons que la classification avec recouvrement est plus appropriée dans la modélisation des relations entre les données et ce pour de nombreuses applications réelles.

Dans certains domaines d'applications, ce type de schéma de classification est naturel. Par exemple en recherche d'information, un document (*texte, image, etc.*) peut aborder plusieurs thématiques ou appartenir à plusieurs domaines différents, encore en biologie, un gène peut influencer plusieurs aspects du métabolisme [SBK03]. Par conséquent, l'utilisation des techniques de classification non supervisée classiques dans chacun de ces domaines peut entraîner la perte d'une information potentiellement utile à l'utilisateur puisque chaque élément ne doit appartenir qu'à une seule classe.

En raison de son importance, la classification avec recouvrement a récemment reçu beaucoup d'attention. Dans ce chapitre, nous nous intéressons à l'ajustement des classes entre elles et plus particulièrement aux recouvrements (*appelés aussi intersections ou empiétements*) entre les classes de mesures. L'intérêt d'avoir recours à la classification à recouvrement est donc d'avoir une nouvelle structure des mesures nous permettant de connaitre les mesures d'intérêt qui peuvent appartenir à plusieurs classes à la fois.

Différentes approches de classification aboutissant à des classes recouvrantes existent dans la littérature, telles que :

- **Algorithme k-moyennes axiales** [Lel93] : cette méthode est une extension des *k-moyennes* (*cf. annexe C, page 207*) pour la recherche de classes recouvrantes. Elle est donc basée sur le principe de classification par centres mobiles, celui des *k-moyennes*. La technique des *k-moyennes axiales* permet d'appliquer sur chaque classe, une analyse factorielle, où les classes seront représentées par des vecteurs normalisés. Ainsi, des

axes de classes sont obtenus, sur lesquels seront projetés des éléments. Cependant, au lieu d'affecter l'élément à la seule classe où sa valeur (*sa distance au centre ou sa probabilité d'appartenance à cette classe*) est la plus grande, il s'agit de l'affecter également à d'autres classes pour lesquelles cette valeur dépasse un certain seuil. Cette méthode est paramétrée par le nombre maximal de classes désiré (k), par une métrique et par le seuil d'appartenance des éléments aux classes qui permet de construire des classes recouvrantes. Cette méthode consomme très peu de mémoire et elle est non déterministe puisque le résultat dépendra de l'initialisation ainsi que de l'ordre d'entrée des données.

- **Algorithme C-moyennes floue** (CMF) [Dun73], [Bez81] : c'est un algorithme de classification floue qui permet à un élément d'appartenir à deux ou plusieurs classes, et ce avec une certaine probabilité. Chaque classe est représentée par son centre de gravité. Pour son exécution, l'algorithme nécessite de connaître le nombre de classes au préalable et génère les classes par un processus itératif en minimisant une fonction objectif. Ainsi, il permet d'obtenir une partition floue des individus en donnant à chaque élément un degré d'appartenance à une classe donnée. Comme les résultats de obtenus par l'algorithme du CMF dépendent de l'étape d'initialisation, (*e.g., l'algorithme des k-moyennes axiales*), il est donc aussi un algorithme non déterministe qui nécessite d'être lancé plusieurs fois (*à chaque fois, un nouveau centre de gravité sera désigné*).
- **Classification pyramidale** [Did86] : cette technique est une extension de la classification hiérarchique qui permet, à partir d'un ensemble fini d'individus et un indice d'agrégation, d'organiser ces individus en une structure de synthèse appelée *pyramide*. Les pyramides constituent une généralisation des hiérarchies en permettant la représentation des recouvrements emboités au lieu de partitions. La classification pyramidale conduit à la structuration des données d'entrée en classes, où chaque classe s'intersecte avec au plus deux autres classes. L'une des propriétés importantes des pyramides pour l'analyse des données est de pouvoir associer un ordre partiel à chaque pyramide [Pat10]. Cette notion d'ordre fait recourt à la notion de dissimilarité Robinsonienne. Le nombre de classes obtenu par la classification pyramidale dépend du niveau de coupure du dendrogramme.
- **Algorithme PoBOC** (*Pole-Based Overlapping Clustering*) [CMV04] : cet algorithme s'appuie sur la notion de pôle, où il s'agit de rechercher dans les ensembles d'individus des zones homogènes formées de plusieurs individus, et situées plutôt en périphérie. La recherche des pôles s'effectue sur un graphe des similarités (*les noeuds du graphe représentent les individus*). Quatre étapes sont essentielles pour l'exécution de cet algorithme : (1) la recherche de pôles, (2) la construction d'une matrice d'appartenance de chaque individu à chacun des pôles, (3) l'affectation des individus à un ou plusieurs pôles, et (4) la construction d'un arbre hiérarchique. Pour ce faire, l'algorithme PoBOC prend en entrée

une matrice de similarité, à partir de laquelle il génère un ensemble de classes non-disjointes présentées sous forme d'une hiérarchie de concepts. Chaque objet ou individu peut donc appartenir à un ou plusieurs concepts. PoBOC est fortement lié aux données textuelles.

– **Analyse factorielle booléenne** [BV10] : c'est une approche de factorisation de données binaires, basée sur la décomposition de matrices binaires en un nombre optimal de facteurs (*ou catégories*). Cette méthode s'appuie sur l'analyse des concepts formels pour découvrir le plus petit nombre de facteurs couvrant la matrice d'entrée, où les concepts formels représentent les facteurs et ces derniers sont donc considérés comme étant des classes. Les classes peuvent être recouvrantes, ayant une signification, et qui sont facilement compréhensibles. Cette méthode est applicable sur des données booléennes et elle est convenable pour une représentation des données dans un espace de faible dimension. Néanmoins, elle nécessite une amélioration en termes de complexité en temps pour analyser des ensembles de données très importants, tout en préservant la nature des données.

Après avoir introduit différentes techniques de classification avec recouvrement, nous cherchons dans ce qui suit à catégoriser l'ensemble des mesures d'intérêt rappelées dans l'*annexe A*, au moyen de méthode(s) de classification empiétante(s). Le choix de la procédure de classification fait l'objet de notre prochaine section.

4.2.2 Choix de la procédure de classification

Étant données les méthodes de classification à recouvrement citées ci-dessus, nous cherchons dans cette section à choisir celle(s) qui soit(ent) la(les) mieux adaptée(s) à nos données, capable(s) de présenter un bon recouvrement des mesures d'intérêt de règles d'association en classes d'objets similaires. Dès lors, nous étudions tout d'abord les limites de chaque méthode afin d'éviter un choix biaisé.

Ainsi, nous trouvons que la première et la quatrième méthode, désignant respectivement l'algorithme des *k-moyennes axiales* et l'algorithme plus général *PoBOC*, sont plutôt motivées par l'application aux données textuelles (*mots ou documents*). Le principe de ces algorithmes est de rechercher, en une ou plusieurs itérations, des centres auxquels sont affectés les objets. Ces centres peuvent être ou bien des axes, tel est le cas de la méthode *k-moyennes axiales*, ou bien des petits ensembles d'objets appelés Pôle dans le cas de la méthode *POBOC*. De plus, la méthode des *k-moyennes axiales* est non déterministe et nécessite la saisie de certains paramètres, e.g., le nombre de classes. De même pour la troisième méthode, *la classification pyramidale*, qui pour fonctionner nécessite un indice de dissimilarité particulier. Cette méthode

dégage aussi certaines faiblesses lors de la classification, puisqu'elle n'autorise pas l'empiétement de plus que deux classes à la fois, ce qui peut être un handicap pour les mesures qui appartiennent à plus que deux classes.

La deuxième méthode, *C-moyennes floue* est non déterministe nécessitant d'être exécutée plusieurs fois afin de détecter les classes fortes (*initialement, le centre de gravité change pour chaque nouvelle exécution*). Cette méthode est, comme celle des *k-moyennes axiales*, basées sur la technique des *k-moyennes*, déjà appliquée dans le *chapitre 3* pour découvrir des classes fortes disjointes (*cf. section classes fortes, page 87*). De même que pour les autres algorithmes, celui des *C-moyennes floue* nécessite des paramètres d'entrée, qui sont le nombre de classes et un indice flou pour fonctionner.

Dans un ancien travail de Suzuki [Suz08], l'auteur a étudié les biais liés à la classification, dont celui des données, où plusieurs paramètres doivent être saisis par l'utilisateur rendant les résultats de la classification moins robustes. Dès lors, nous considérons cette étude durant notre recherche de la bonne technique de classification, capable de nous fournir le meilleur recouvrement des mesures.

Toutefois, nous remarquons que la cinquième méthode, celle de l'*analyse factorielle booléenne* [BV10] ne présente pas les limitations des autres techniques de classification avec recouvrement retenues : comme par exemple la saisie de paramètres d'entrée, tel que le nombre de classes. Ce nombre est plutôt identifié par le nombre minimal de facteurs couvrant l'ensemble des données. L'AFB a été plutôt critiquée en termes de complexité en temps, mais cela ne nous concerne pas puisque nos données d'entrée ne sont pas très importantes. D'où notre motivation pour appliquer cette méthode sur nos données d'entrée qui va nous révéler un bon recouvrement des mesures d'intérêt. Selon les auteurs [BV10], cette méthode est aussi la plus performante pour la décomposition de matrices binaires, avec un nombre optimal de facteurs couvrant les données. Une autre raison qui nous a motivé à utiliser cette technique pour catégoriser les mesures, est l'application d'une nouvelle méthode, qui est basée sur des notions différentes que celles que nous avons utilisées dans la précédente classification serait intéressant.

Dans ce qui suit, nous proposons la méthodologie de travail qui sera suivie tout au long de ce chapitre.

4.2.3 Mise en oeuvre de la classification

Nous résumons dans ce qui suit les caractéristiques de la procédure de classification avec recouvrement utilisée. En outre, nous présentons un résumé de notre procédure d'analyse du comportement des mesures d'intérêt.

Caractéristiques de la classification Des choix de critères variés sont laissés à notre initiative, tels que :

- le choix des variables et de leurs modalités : ***matrice mesure-propriété*** ;
- le format des variables : ***variables booléennes*** ;
- Choix de la méthode de classification avec recouvrement : ***analyse factorielle booléenne***.

Étant donnés ces critères de classification, nous pouvons maintenant présenter notre méthodologie de travail.

Processus d'analyse La procédure d'analyse du comportement des mesures d'intérêt est résumée dans ce qui suit :

Étape 1 : *Préparation des données*

Cette première étape consiste à traiter les données d'entrée, recueillies dans le cadre de l'étude des mesures d'intérêt selon les propriétés, pour les adapter à la méthode de classification avec recouvrement que nous allons utiliser. Ayant des données qualitatives nominales, nous appliquons un codage disjonctif complet sur notre matrice de mesures-propriétés pour obtenir une nouvelle avec des données binaires.

Étape 2 : *Application de méthode de classification avec recouvrement*

L'étape 2 consiste à appliquer l'analyse factorielle booléenne sur la matrice binaire obtenue dans l'*étape 1*. Cette méthode utilise les concepts formels pour donner une nouvelle structure aux données binaires sous forme de facteurs ou de catégories de mesures.

Étape 3 : *Interprétation*

L'étape 3 consiste à interpréter les catégories de mesures empiétantes identifiées, afin de comprendre le comportement des mesures d'une même classe.

Étape 4 : *Comparaison*

Finalement, l'étape 4 consiste à comparer les résultats de la classification avec recouvrement, identifiés dans l'*étape 2*, avec ceux obtenus dans le *chapitre 3* par les méthodes de classification non supervisée, *CAH* et *k-moyennes*.

Étape 5 : *Discussion*

Cette dernière étape consiste à confronter les deux résultats (*ce sont les résultats obtenus par les méthodes sans et avec recouvrement*) de la classification pour rechercher des groupes stables de mesures, i.e., des ensembles de mesures que nous retrouvons toujours ensemble quelque soit la technique de classification utilisée.

Après avoir présenté les différentes étapes que nous allons suivre tout au long de ce chapitre, nous procédons dans la section suivante à la définition de l'analyse factorielle booléenne.

4.3 Analyse factorielle booléenne (AFB)

Soit I une matrice booléenne (*binaire*) de dimension $m \times c$. L'objectif de l'analyse factorielle booléenne (AFB) [BV10], aussi appelée analyse factorielle de données binaires (*booléennes*), est de trouver une décomposition de la forme :

$$I = X \circ Y \tag{4.1}$$

de I en deux matrices booléennes X et Y de dimensions respectives $m \times k$ et $k \times c$, avec $\circ$ désignant le produit booléen des deux matrices, i.e.,

$$(X \circ Y)_{ij} = \max_{l=1}^{k} \min(X_{il}, Y_{lj}).$$

La dimension intérieure, k, de la décomposition peut être interprétée comme étant le nombre de facteurs qui peuvent être utilisés pour décrire les données d'origine. À savoir, $X_{il} = 1$ si et seulement si le $l$ième facteur s'applique au $i$ème objet et $Y_{lj} = 1$ si et seulement si le $j$ème attribut est l'une des manifestations du $l$ième facteur. Le modèle factoriel représentant (4.1) possède donc la signification suivante : l'objet i possède l'attribut j si et seulement si, il existe un facteur l qui s'applique à i et pour qui j est l'une des manifestations particulières de l. Cependant, bien que l'objectif général de l'analyse factorielle classique (*AFC*) et celui de l'analyse factorielle booléenne soit le même, l'AFB reste substantiellement différente de l'AFC. Cette différence est due à la nature des données (*données booléennes*) et l'algèbre s'inspirant derrière l'AFB.

Dans [BV10], les auteurs ont présenté une méthode permettant la découverte de décompositions (4.1) où le nombre de facteurs k est aussi petit que possible. La méthode utilise les concepts formels du contexte formel $\langle O, A, I \rangle$, où $O = \{1, \ldots, m\}$ et $A = \{1, \ldots, c\}$ sont des ensembles et $I \subseteq O \times A$. Les éléments de O sont appelés les objets et ceux de A les attributs, ils correspondent respectivement aux lignes et colonnes de I. L'ensemble de couple I est considéré comme une relation et est donc noté mIc au lieu de $(m, c) \in I$ ce qui se dit : "*l'objet m possède l'attribut c*". Rappelons qu'un concept formel de $\langle O, A, I \rangle$ est une paire $\langle C, G \rangle$ d'ensembles $C \subseteq O$ et $G \subseteq A$, tel que G représente l'ensemble des attributs communs à tous les objets de C et, inversement, C représente l'ensemble de tous les objets qui ont tous les attributs de G.

Nous mentionnons également que l'ensemble de tous les concepts formels de $\langle O, A, I\rangle$ muni d'une hiérarchie de la forme sous-concept-super-concept, est appelé le treillis de concepts de $\langle O, A, I\rangle$ et est noté par $\mathcal{B}(O, A, I)$ [GW97]. Soit

$$\mathcal{F} = \{\langle C_1, G_1\rangle, \ldots, \langle C_k, G_k\rangle\} \tag{4.2}$$

un ensemble de concepts formels de $\langle O, A, I\rangle$. Considérons la matrice booléenne $X_{\mathcal{F}}$, de dimension $m \times k$ et la matrice booléenne $Y_{\mathcal{F}}$, de dimension $k \times c$, définies par

$$(X_{\mathcal{F}})_{il} = 1 \text{ ssi } i \in C_l \quad \text{et} \quad (Y_{\mathcal{F}})_{lj} = 1 \text{ ssi } j \in G_l. \tag{4.3}$$

Notons par $\rho(I)$ le plus petit nombre k, nommé le rang de Schein de I, pour une décomposition de I avec k facteurs. Le théorème suivant [BV10] montre que l'utilisation de concepts formels comme étant des facteurs, de même que dans (4.3), est optimale puisqu'elle nous permet d'atteindre le rang Schein.

Théorème 1 *Pour chaque matrice booléenne I, il existe $\mathcal{F} \subseteq \mathcal{B}(O, A, I)$ tels que $I = X_{\mathcal{F}} \circ Y_{\mathcal{F}}$ et $|\mathcal{F}| = \rho(I)$.*

Exemple 1 *À titre d'illustration, considérons la matrice booléenne I de dimension 4×5 représentée ci-dessous. En utilisant la méthode AFB décrite précédemment (où les concepts formels représentent les facteurs $\mathcal{F}$), nous pouvons trouver une décomposition de la matrice en des matrices booléennes $X_{\mathcal{F}}$ et $Y_{\mathcal{F}}$, de dimensions respectives 4×3 et 3×5, où le nombre de facteurs k (et aussi le rang de Schein $\rho(I)$), est égal à 3 :*

$$\begin{array}{ccccc} & I & & = & X_{\mathcal{F}} & \circ & Y_{\mathcal{F}} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} & & & = & \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} & \circ & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Considérons la matrice booléenne d'entrée comme étant un contexte formel $\langle O, A, I\rangle$, où $O = \{1, \ldots, 4\}$ et $A = \{1, \ldots, 5\}$, les trois concepts formels suivants de $\langle O, A, I\rangle$ sont utilisés comme des facteurs :

$$\mathcal{F} = \{\langle\{1,2,3\}, \{1,2\}\rangle, \langle\{2,4\}, \{1,5\}\rangle, \langle\{3\}, \{1,2,3,4\}\rangle\}.$$

Comme il a été démontré dans [BV10], il serait utile de considérer les concepts formels comme des facteurs puisque les concepts formels peuvent être facilement interprétés. En effet, chaque facteur, i.e., un concept formel $\langle C_l, G_l\rangle$, se compose d'un ensemble C_l d'objets (*les objets représentent les mesures d'intérêt dans notre cas*) et un ensemble G_l d'attributs (*les attributs représentent les propriétés des mesures dans notre cas*). L'ensemble C_l ne contient

que les objets auxquels s'appliquent tous les attributs de G_l et l'ensemble G_l contient tous les attributs partagés par tous les objets de C_l.

D'un point de vue classification, les facteurs $\langle C_l, G_l \rangle$ peuvent donc être considérés comme des classes C_l, décrites par des attributs de G_l ayant une signification naturelle, facilement compréhensible. Comme le problème du calcul du plus petit ensemble de facteurs est NP-difficile, un algorithme d'approximation a été proposé dans [BV10, Algorithme 2]. Cet algorithme sera utilisé dans ce qui suit (*et a été aussi utilisé dans l'exemple ci-dessus*). Il est basé sur une stratégie glouton qui permet de sélectionner à chaque étape, de larges concepts formels couvrant la plupart des données non encore couvertes. L'algorithme utilise une méthode heuristique de sélection de concepts formels qui permet d'éviter le calcul de l'ensemble du treillis de concepts, rendant ainsi la sélection à la fois beaucoup plus rapide et avec une qualité comparable par rapport à une stratégie gourmande directe effectuant une recherche exhaustive, ce qui est représenté par [BV10, Algorithme 1].

Dans la prochaine section, nous procédons au cadre applicatif par la classification des mesures d'intérêt selon la méthode d'analyse factorielle booléenne que nous venons de définir.

4.4 Classification des mesures d'intérêt au moyen de l'AFB

Dans cette section, nous présentons une nouvelle classification des mesures d'intérêt en utilisant la méthode AFB. Pour ce faire, nous avons en entrée une matrice booléenne I, où les lignes correspondent aux mesures d'intérêt et les colonnes à leurs propriétés. En utilisant *l'algorithme 2* de [BV10], nous calculons un ensemble de facteurs $\mathcal{F}$ (*i.e. des concepts formels*), qui décomposent I, i.e., pour lesquels $I = X_{\mathcal{F}} \circ Y_{\mathcal{F}}$. Les concepts formels $\langle C_i, G_i \rangle \in \mathcal{F}$, (*cf. section 4.2*), nous fournissent des classes de la façon suivante ; chaque concept formel $\langle C_i, G_i \rangle$ se compose d'un ensemble de mesures C_i et d'un ensemble de leurs propriétés G_i. C_i peut être considéré comme un ensemble de mesures et G_i comme étant sa description. Dans ce qui suit, on note les facteurs $\langle C_i, G_i \rangle$ par F_i. Pour plus de commodité, nous identifions parfois l'ensemble C_i (*i.e., les mesures couvertes par* F_i), avec F_i. Autrement dit, au lieu d'utiliser la notation $C_i \subseteq C$ qui est la mieux appropriée, nous considérons plutôt $F_i \subseteq C$ pour indiquer que les mesures couvertes par le facteur F_i sont incluses dans un certain ensemble C. Nous mentionnons également que, selon la nature de l'AFB, les premiers facteurs représentent généralement le plus de données et sont donc les plus intéressants, tandis que les derniers peuvent être ignorés. Par conséquent, nous n'examinons que les premiers facteurs considérés comme étant des candidats pour les classes de mesures.

Nous cherchons maintenant à appliquer l'AFB sur la matrice d'évaluation des mesures (*cf. chapitre 2, pages 70 et 71*) et à comparer ensuite les groupes de mesures obtenus par cette

technique avec ceux identifiés par les deux méthodes de classification sans recouvrement (*k-moyennes et CAH*).

4.4.1 Entrée : Mesures et leurs propriétés

Dans cette section, nous analysons les mesures d'intérêt et leurs propriétés. Pour ce faire, nous reprenons l'étude que nous avons réalisée dans le *chapitre 2*, où nous avons proposé une évaluation de 61 mesures d'intérêt selon 19 propriétés, à savoir P_3 à P_{21}. Cette évaluation se traduit par une matrice 61×19, utilisée pour regrouper les mesures en fonction de leurs propriétés.

La matrice de mesures selon les propriétés que nous utilisons pour la classification des mesures au moyen de l'AFB est illustrée dans la *table 4.1*. Elle diffère de celle décrite dans la *table 2.3, pages 70 et 71* au niveau de l'évaluation de la propriété P_{14} (*Tolérance aux premiers contre-exemples*). Pour cette propriété, nous retenons ses 3 propriétés binaires : $P_{14.1}$ si la mesure est concave, $P_{14.2}$ si la mesure est linéaire, et $P_{14.3}$ si la mesure est convexe (*cf. sous-section 2.3.14, page 64 du chapitre 2*), et nous obtenons une matrice composée de 61 mesures et décrite par 21 propriétés (*18 propriétés binaires et une propriété à trois valeurs, à savoir* P_{14}).

Ayant précisé nos données d'entrée, nous procédons maintenant à la classification des mesures en appliquant l'AFB.

4.4.2 Sortie : Classification utilisant les facteurs booléens

Dans cette étape, nous calculons tout d'abord la décomposition de la matrice de la *table 4.1* en utilisant l'*algorithme 2* de [BV10]. Ce calcul nous permet d'obtenir 28 facteurs. Le dernier facteur, le $28^{ème}$, couvre les dernières colonnes de la matrice d'entrée contenant les '*1s*' qui n'étaient pas identifiés par les facteurs précédents, désignant ainsi la condition d'arrêt de l'algorithme. Cependant, puisque les facteurs sont triés du plus au moins important, et que leur importance est déterminée par le nombre de '*1s*' dans la matrice d'entrée couverts par le facteur en question, nous étendons à présent la matrice booléenne originale (*table 4.1*) de dimension 61×21 en ajoutant pour chaque propriété sa négation. Un codage disjonctif complet est alors appliqué sur la matrice d'origine afin de pallier la principale limite de l'AFB, qui comme nous l'avons décrit, est basée sur l'analyse des concepts formels. Elle ne traite donc que le cas '1', mais pas le cas '0'. D'où le besoin de doubler les descripteurs, par l'application d'un codage disjonctif complet, qui va permettre de tenir compte aussi des cas '0'. Une nouvelle matrice booléenne de dimension 61×42 est alors construite. À partir de cette matrice, nous obtenons 38 facteurs, notés $F_1, \ldots, F_{38}$. Ce nombre est bien évidemment plus élevé que celui discerné par les premières expérimentations réalisés sur la matrice de dimension 61×21, puisque le

	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14.1	P15	P16	P17	P18	P19	P20	P21	P14.2	P14.3
Corrélation		1	1	1	1	1	1			1	1			1	1	1			1	1	
Cohen		1	1	1	1	1	1			1	1					1			1	1	
Confiance	1	1	1	1				1	1										1	1	
Confidence causale	1	1	1	1	1	1		1											1	1	
Pavillon	1	1		1	1	1	1			1	1				1				1	1	
Ganascia	1	1	1	1				1	1						1				1	1	
Confirmation causale	1	1	1	1	1	1													1	1	
Confirmation descriptive	1	1		1					1						1				1	1	
Conviction	1	1	1	1	1	1	1			1	1								1		1
Cosinus		1		1		1													1	1	
Couverture	1																			1	
Dépendance	1						1												1	1	
Dépendance Causale	1	1		1	1	1													1	1	
Gray Orlowska	1	1			1														1		
Facteur bayésien	1	1		1	1	1	1			1	1		1						1		1
Loevinger	1	1	1	1	1	1	1	1		1	1								1	1	
Force Collective		1	1	1	1	1	1			1	1					1			1		1
Fukuda	1	1		1														1	1	1	
Gain Informationnel		1		1	1	1	1			1	1	1							1		
Goodman		1	1	1	1		1	1		1	1			1	1	1			1	1	
Indice d'implication	1	1	1				1											1	1	1	
IPEE	1	1	1	1					1			1					1	1			
IP3E	1	1	1	1					1			1					1	1	1		
IPD		1	1	1		1						1	1				1	1	1		
II	1	1	1	1	1	1	1			1	1	1	1				1	1			
IIE	1	1	1	1	1							1					1	1	1		
IIER	1	1	1	1	1		1			1	1	1					1	1	1		
IVL		1	1	1	1	1	1			1	1	1					1	1			
Intérêt		1		1	1	1	1			1	1								1	1	
Jaccard		1		1		1													1		1
J-mesure	1						1			1									1		1
Klosgen	1	1			1		1			1	1								1		1
Laplace	1	1	1	1					1										1	1	
Mgk	1	1	1	1	1		1	1		1	1				1				1	1	
Moindre contradiction	1	1		1					1										1	1	
Pearl			1				1			1						1			1	1	
Piatetsky-Shapiro		1	1	1	1	1	1			1	1			1	1	1		1	1	1	
Précision		1	1	1	1	1										1			1	1	
Prévalence	1	1																		1	
Q de Yule		1	1	1	1		1	1		1	1	1	1	1	1	1			1		
Rappel	1	1		1		1													1	1	
Gini	1															1			1		1
Risque relatif	1	1		1	1	1	1			1	1								1		1
Sebag	1	1		1					1										1		1
Support		1		1															1	1	
Support sens unique	1	1			1	1	1			1	1								1		1
Support double sens		1			1	1	1			1	1								1		1
Taux d'exemples	1	1	1	1				1	1			1									
VT100		1	1	1	1	1								1	1	1	1		1	1	
Variation support			1				1			1						1			1		1
Y de Yule		1	1	1	1		1	1		1	1		1	1	1	1			1		1
Zhang	1	1	1	1	1		1	1		1	1	1	1		1				1		
Confiance-confirmée causale	1	1	1	1	1	1		1											1	1	
Czekanowski-dice		1		1		1													1	1	
Fiabilité négative	1	1	1	1	1	1		1											1	1	
Mutual information	1															1			1		1
Kulczynski		1		1		1													1		1
Leverage	1	1		1	1	1													1	1	
Nouveauté		1	1	1	1	1	1			1	1			1	1	1			1	1	
Ratio des chances		1	1	1	1	1	1			1	1					1			1		1
Spécificité	1	1		1	1	1													1	1	

TABLE *4.1: Matrice booléenne d'entrée décrivant les mesures d'intérêt par leurs propriétés.*

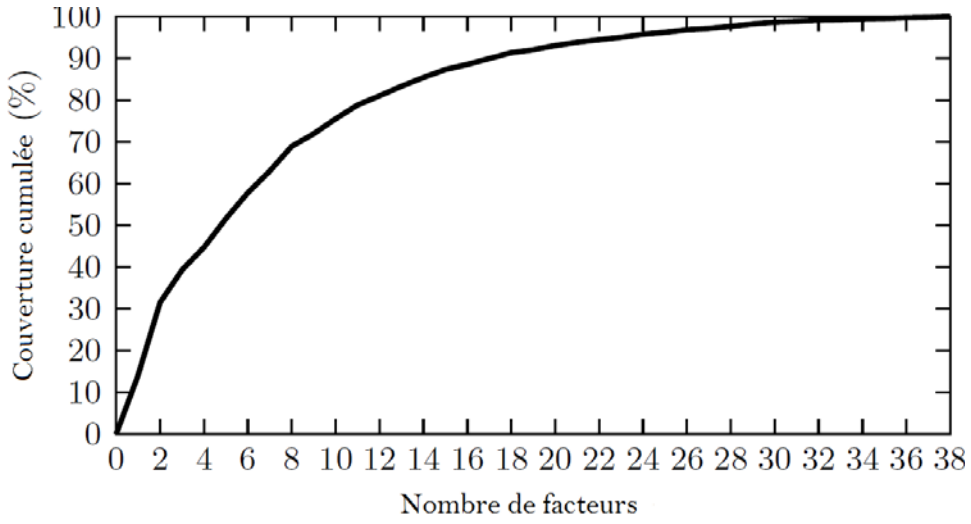

FIGURE *4.1: Couverture cumulée de la matrice d'entrée de la table 4.1, étendue par des propriétés inversées et par les facteurs, obtenus par la décomposition de la matrice.*

nombre d'attributs de la nouvelle matrice a doublé (*nous passons de 21 attributs à 42*).

Les 38 facteurs sont présentés dans les *tables 4.2* et *4.3*. La *table 4.2* représente la matrice objet-facteur décrivant les mesures d'intérêt par les facteurs, *i.e.* la matrice $X_{\mathcal{F}}$ (*voir la section 4.3*). La *table 4.3* représente la matrice facteurs-propriétés expliquant les facteurs par les propriétés des mesures, *i.e.* la matrice $Y_{\mathcal{F}}$ (*voir la section 4.3*). Plusieurs de ces facteurs peuvent être écartés parce qu'ils ne sont pas très importants. Le diagramme de la *figure 4.1* présente la couverture cumulée de la matrice de dimension 61×42 par les 38 facteurs.

Les facteurs ainsi obtenus par l'AFB doivent être analysés dans la recherche de ceux qui sont les plus intéressants. Une fois identifiés, nous reprenons les résultats de la classification effectuée dans le *chapitre 3* en utilisant la méthode ascendante hiérarchique et la méthode des *k-moyennes*, et qui a révélé 7 classes de mesures, pour les comparer avec les facteurs retenus de l'AFB. Cette comparaison se fait au fur et à mesure que nous interprétons les facteurs. Ainsi, les deux *étapes 3 et 4* de notre processus d'analyse seront réalisées en parallèle.

4.4.3 Interprétation et comparaison

Comme il a été mentionné précédemment, 38 facteurs ont été identifiés. Les 21 premiers facteurs d'entre eux couvrent 94% de la matrice d'entrée des mesures-propriétés (*1s dans la matrice*), les neuf premiers couvrent 72%, et les cinq premiers facteurs couvrent 52%. Nous remarquons également que les dix premiers facteurs couvrent l'ensemble des mesures.

Toutefois, nous rappelons que contrairement à la classification non supervisée classique, qui permettait d'obtenir des groupes disjoints, les facteurs booléens représentent des groupes

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24	F25	F26	F27	F28	F29	F30	F31	F32	F33	F34	F35	F36	F37	F38
Corrélation	1					1			1				1								1					1					1								
Cohen	1					1							1					1			1										1							1
Confiance		1		1								1						1												1								
Confiance causale		1				1				1		1						1												1								
Pavillon	1					1				1										1	1					1					1		1					
Ganascia		1		1								1														1				1								
Confirmation causale		1				1				1	1							1												1								
Confirmation descriptive		1		1										1												1				1			1					
Conviction	1						1								1			1			1																1	
Cosinus		1			1						1		1	1																		1	1					
Couverture			1							1				1										1	1					1								
Dépendance			1							1				1						1				1							1								
Dépendance causale		1			1	1				1	1																			1			1					
Dépendance pondérée			1								1						1												1	1			1					
Facteur bayésien	1				1										1				1		1												1			1		
Loevinger	1					1				1		1						1			1										1								
Force collective	1						1						1					1			1						1									1	1	
Fukuda										1	1						1											1		1			1	1				
Gain informationnel	1												1							1	1	1										1	1					
Goodman	1								1			1						1								1					1								
Indice d'implication			1							1								1	1									1			1								
IPEE				1				1										1				1			1			1	1				1					
IP3E				1				1			1							1				1						1	1				1					
IPD								1			1		1						1			1						1					1					
II								1							1				1		1	1			1								1	1				
IIE								1			1						1	1				1								1			1	1				
IIER								1							1		1	1				1											1	1				
IVL								1					1					1			1	1			1								1	1				
Intérêt	1				1	1							1							1	1									1	1	1					
Jaccard					1		1				1		1	1																		1	1					
J-mesure			1				1													1				1											1		
Klosgen	1						1								1		1			1								1					1			1		
Laplace		1		1							1							1												1								
MGK	1									1		1					1									1					1							
Moindre contradiction		1		1							1			1																1			1					
Pearl			1													1		1													1							1
Piatetsky-Shapiro						1			1				1								1					1					1			1				
Précision						1					1		1					1					1															1
Prévalence			1							1				1											1					1			1					
Q de Yule	1								1			1					1		1			1													1			
Rappel		1			1					1	1			1																1			1					
Gini			1				1				1			1									1	1			1											1
Risque relatif	1				1		1								1					1	1												1			1		
Sebag				1			1				1			1																			1					
Support		1									1			1				1															1	1					
Support sens unique	1				1		1								1					1								1					1			1		
Support double sens	1				1		1													1								1				1	1			1		
Taux d'exemples				1								1						1				1			1					1								
VT100						1			1				1										1												1				
Variation support			1				1									1		1																		1	1	
Y de Yule	1								1								1		1																	1	1	
Zhang	1											1			1		1		1			1														1		
Confiance-confirmée causale		1				1				1		1						1												1								
Czekanowski-dice		1			1						1		1	1																		1	1					
Fiabilité négative		1				1				1		1						1												1								
Information mutuelle			1				1				1			1									1	1			1											1
Kulczynski					1		1				1		1	1																		1	1					
Leverage		1			1	1				1	1																			1			1					
Nouveauté	1					1			1				1								1						1				1							
Ratio des chances	1						1						1					1			1						1										1	1
Spécificité		1			1	1					1	1																		1			1					

TABLE *4.2: Les mesures d'intérêt décrites par les facteurs suite à la décomposition de la matrice d'entrée de la table 4.1, étendue par les propriétés inversées.*

	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14.1	P15	P16	P17	P18	P19	P20	P21	P14.2	P14.3	$\overline{P3}$	$\overline{P4}$	$\overline{P5}$	$\overline{P6}$	$\overline{P7}$	$\overline{P8}$	$\overline{P9}$	$\overline{P10}$	$\overline{P11}$	$\overline{P12}$	$\overline{P13}$	$\overline{P14.1}$	$\overline{P15}$	$\overline{P16}$	$\overline{P17}$	$\overline{P18}$	$\overline{P19}$	$\overline{P20}$	$\overline{P21}$	$\overline{P14.2}$	$\overline{P14.3}$
F1		1			1		1			1	1									1										1								1	1			
F2		1		1																1	1							1			1	1	1	1	1		1	1	1			1
F3																									1		1		1	1		1	1	1	1	1			1				
F4	1	1		1					1																	1	1	1			1	1		1	1		1					
F5		1				1														1				1					1	1			1		1	1	1	1	1			
F6		1		1	1	1														1	1								1				1	1								1
F7																				1		1							1				1	1	1	1		1	1		1	
F8		1	1	1								1					1	1											1						1	1	1				1	1
F9		1	1	1	1									1	1	1				1		1							1													
F10	1																				1								1				1	1	1		1	1				1
F11																				1							1	1		1	1			1	1							
F12		1	1	1				1																														1	1			1
F13		1		1		1																1							1	1												
F14																								1		1			1		1	1	1	1	1			1	1			
F15	1	1			1		1			1	1																		1					1		1				1		
F16			1				1			1						1				1		1	1		1	1	1		1	1	1	1	1	1			1	1				
F17		1																		1							1		1													
F18			1																												1	1	1									
F19		1		1									1																1											1			
F20							1													1				1					1				1	1		1	1	1				
F21		1		1	1	1	1			1	1																		1													
F22		1		1								1																												1	1	
F23																1				1							1	1	1	1	1	1	1					1				
F24	1																						1	1	1	1	1		1	1		1	1	1	1	1		1	1			
F25																																		1	1	1				1		1
F26		1		1											1					1	1												1	1				1				1
F27																1				1		1							1				1					1	1		1	
F28		1																1							1			1			1	1	1								1
F29		1			1															1				1	1				1	1	1	1	1	1	1			1			
F30	1																										1			1	1		1	1		1					1
F31							1													1	1								1				1	1				1				1
F32		1																		1			1						1	1			1	1	1	1	1	1				
F33		1																					1						1					1		1	1					
F34		1		1														1											1													1
F35		1	1	1	1												1												1	1												1
F36		1	1	1	1		1	1		1	1		1		1					1							1		1									1	1		1	
F37							1			1										1		1							1				1					1	1		1	
F38																1				1									1	1			1	1	1	1		1	1			

TABLE *4.3: Matrice facteur-propriété obtenue par la décomposition de la matrice d'entrée de la table 4.1, étendue par des propriétés inversées. Les facteurs sont décrits en terme de propriétés originales et inversées.*

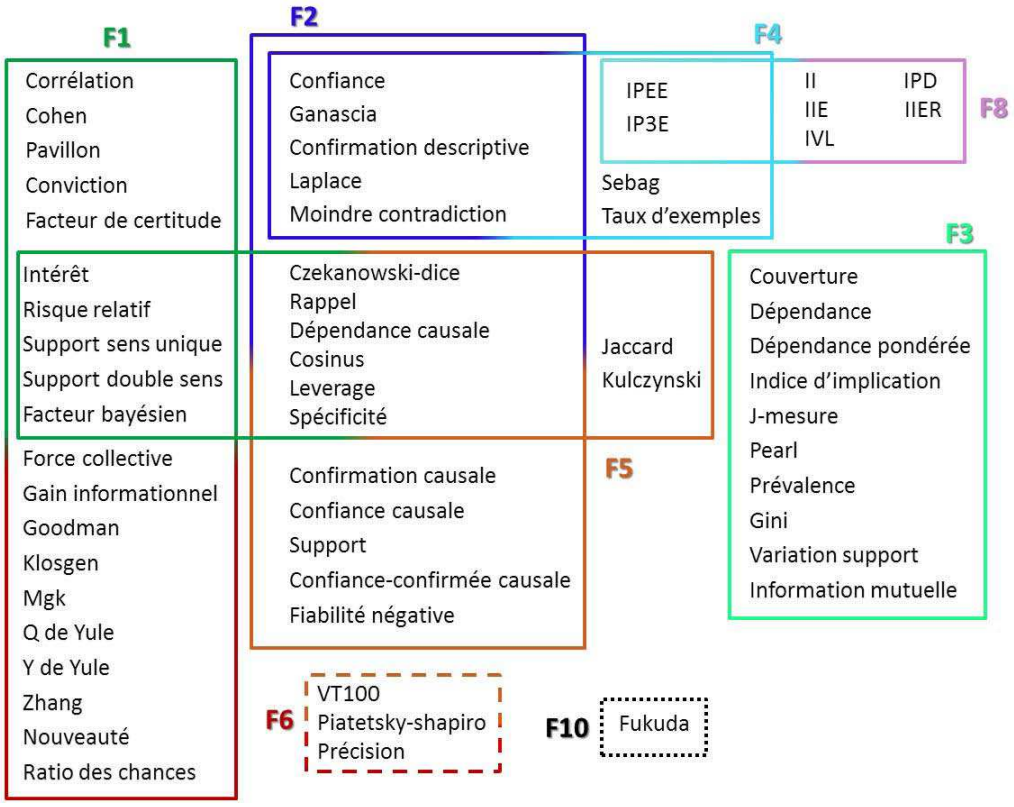

FIGURE *4.2: Diagramme de Venn des facteurs sélectionnés de la table 4.3.*

avec recouvrement. À savoir, les groupes qui sont représentés dans la *figure 4.2* décrivent le diagramme de Venn des cinq premiers facteurs booléens, plus le huitième et une partie du sixième ainsi que le dixième pour couvrir l'ensemble des mesures. En regardant la *table 4.2*, nous remarquons que les facteurs F_7 et F_9 ne couvrent pas les mesures qui manquaient aux premiers facteurs et donc nous les ignorons dans le diagramme de Venn.

Nous reprenons les 7 classes de mesures C_1 à C_7, révélés dans le *chapitre 3* suite à un consensus sur la classification obtenu par les deux méthodes de classification : *hiérarchique* et *k-moyennes*. Étant données ces 7 classes, il est possible maintenant de les comparer avec les facteurs révélés par la méthode AFB comme suit.

Facteur 1. Le premier facteur F_1 s'applique à 20 mesures, voir la *table 4.2* : *Corrélation, Cohen, Pavillon, Conviction, Facteur bayésien, Loevinger, Force collective, Gain informationnel, Goodman, Intérêt, Klosgen, M_{GK}, Q de Yule, Risque relatif, Support sens unique, Support double sens, Y de Yule, Zhang, Nouveauté, et Ratio des chances*. Ces mesures partagent les 9 propriétés suivantes : P_4, P_7, P_9, $\overline{P_{11}}$, P_{12}, P_{13}, $\overline{P_{19}}$, $\overline{P_{20}}$, P_{21}, (voir la *table 4.3*).

Interprétation. Le facteur s'applique à des mesures qui croissent en fonction du nombre d'exemples et qui possèdent une valeur fixe dans le cas de l'indépendance (*ce qui permet d'identifier la zone attractive et répulsive d'une règle*). Le facteur s'applique également à des mesures uniquement descriptives, discriminantes et qui ne sont pas basées sur un modèle probabiliste.

Comparaison. Lorsque l'on regarde les résultats de la classification figurant dans le chapitre précédent (*cf. Classification définitive, page 89*), nous trouvons que F_1 couvre deux classes : C_6 et C_7, dont l'union révèle 15 mesures. Ces classes sont étroitement liées selon le dendogramme obtenu avec la classification ascendante hiérarchique (*illustrée dans le chapitre 3, page 85*).

Les 5 mesures manquantes (*Force collective, Cohen, Corrélation, Nouveauté, et Ratio des chances*) forment la classe G_{p8}, obtenue avec la méthode des *k-moyennes* dans le *chapitre 3* en utilisant la distance euclidienne. Cependant, comme nous l'avons déjà expliqué dans le même chapitre (*cf. Étude des mesures instables, page 107*), que les mesures de G_{p8} sont très proches en termes de distance à certaines mesures des classes C_6 et C_7, ce qui confirme leur couverture par le même facteur F_1.

Facteur 2. F_2 s'applique à 16 mesures, à savoir : *Confiance, Confiance causale, Ganascia, Confirmation causale, Confirmation descriptive, Cosinus, Dépendance causale, Laplace, Moindre contradiction, Rappel, Support, Confiance-confirmée causale, Czekanowski-dice, Fiabilité négative, Leverage et Spécificité*. Ces différentes mesures partagent les *12* propriétés suivantes : P_4, P_6, $\overline{P_9}$, $\overline{P_{12}}$, $\overline{P_{13}}$, $P_{14.2}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{18}}$, $\overline{P_{19}}$, $\overline{P_{20}}$ et P_{21}.

Interprétation. Le facteur s'applique à des mesures qui croissent en fonction du nombre

d'exemples et qui possèdent une valeur variable dans le cas de l'indépendance. Ceci implique que les zones attractives et répulsives d'une règle ne sont pas identifiables. Le facteur s'applique aussi à des mesures uniquement discriminantes, indifférentes aux premiers contre-exemples, et qui ne sont basées sur aucun modèle probabiliste.

Comparaison. F_2 correspond à deux classes, C_4 et C_5, illustrées dans le *chapitre 3*. Le dendogramme (*figurant aussi dans le chapitre précédent, page 85*) confirme ce rapprochement. L'union de ces deux classes $C_4 \cup C_5$ comprend 21 mesures. Certaines (*Jaccard, Kulczynski, Précision, Taux d'exemples, et Sebag*) parmi ces mesures qui sont à la fois présentes dans $C_4 \cup C_5$ et manquantes de F_2 ne sont pas couvertes par F_2 car elles ne sont pas indifférentes aux premiers contre-exemples.

Facteur 3. F_3 s'applique à 10 mesures, à savoir : *Couverture, Dépendance, Dépendance pondérée, Indice d'implication, J-mesure, Pearl, Prévalence, Gini, Variation support, et Information mutuelle*. Ces mesures partagent les 10 propriétés suivantes : $\overline{P_6}$, $\overline{P_8}$, $\overline{P_{10}}$, $\overline{P_{11}}$, $\overline{P_{13}}$, $\overline{P_{14.1}}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{17}}$ et $\overline{P_{19}}$.

Interprétation. Le facteur s'applique à des mesures qui sont décroissantes en fonction du nombre d'exemples et ayant des valeurs variables au niveau de l'équilibre et de l'implication logique. Leur comportement commun les rend unique par rapport aux autres mesures étudiées et explique l'absence d'un facteur qui chevauche avec F_3.

Comparaison. F_3 correspond à la classe C_3 (*cf. Classification définitive, page 89*), qui contient 8 mesures. Les deux mesures, *Variation support* et *Pearl*, appartiennent à C_3 selon la méthode des *k-moyennes* uniquement mais le dendogramme (*voir page 85*) prouve qu'elles sont proches à cette classe. Nous remarquons une forte correspondance entre les résultats obtenus par les facteurs booléens et par les méthodes de classification sans recouvrement pour ce groupe de mesures.

Facteur 4. F_4 s'applique aux 9 mesures suivantes : *Confiance, Ganascia, Confirmation descriptive, IPEE, IP3E, Laplace, Moindre contradiction, Sebag, et Taux d'exemples*. Ces mesures partagent les *12* propriétés suivantes : P_3, P_4, P_6, P_{11}, $\overline{P_7}$, $\overline{P_8}$, $\overline{P_9}$, $\overline{P_{12}}$, $\overline{P_{13}}$, $\overline{P_{15}}$, $\overline{P_{16}}$ et $\overline{P_{18}}$.

Interprétation. Le facteur s'applique à des mesures croissantes en fonction du nombre d'exemples et ayant une valeur fixe dans le cas de l'équilibre.

Comparaison. F_4 s'applique essentiellement aux mesures de la classe C_5, obtenue dans le chapitre précédent. Nous retrouvons les 7 mesures de C_5 dans F_4. Les seules deux mesures manquantes, *IPEE* et *IP3E*, qui appartiennent à une classe différente C_2, sont identifiées également au sein du facteur F_8, qui se chevauche avec F_4. Ceci peut s'expliquer par le fait que *IPEE* et *IP3E* partagent la propriété P_{11} avec d'autres mesures du même facteur F_4.

Facteur 5. F_5 s'applique à 13 mesures, à savoir : *Intérêt, Risque relatif, Support sens unique, Support double sens, Facteur bayésien, Czekanowski-dice, Rappel, Dépendance causale, Cosinus, Leverage, Spécificité, Jaccard et Kulczynski*. L'ensemble de ces mesures partagent les *13* propriétés suivantes : P_4, $\overline{P_5}$, P_8, $\overline{P_{10}}$, $\overline{P_{11}}$, $\overline{P_{14.1}}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{17}}$, $\overline{P_{18}}$, $\overline{P_{19}}$, $\overline{P_{20}}$ et P_{20}.

Interprétation. Le facteur s'applique uniquement à des mesures descriptives et discriminantes, qui décroissent lorsque le nombre de motifs satisfaisant le conséquent croit. Il s'applique également aux mesures non constantes dans les cas de l'équilibre et de l'implication logique et qui ne sont pas basées sur un modèle probabiliste.

Comparaison. Le facteur F_5 couvre principalement les deux classes C_4 et C_7, résultant de la classification sans recouvrement décrite dans le chapitre précédent. $C_4 \cup C_7$ contient 24 mesures. Ainsi, nous avons 11 mesures manquantes de F_5, qui sont les suivantes : *Précision, Support, Confirmation causale, Confiance causale, Confiance-confirmée causale, Fiabilité négative* qui appartiennent à C_4, et *Gain informationnel, Conviction, Loevinger, Pavillon, Klosgen* qui représentent les mesures manquantes de C_7.

Facteur 6. F_6 s'applique à 16 mesures, à savoir : *VT100, Piatetsky-Shapiro, Précision, Corrélation, Cohen, Pavillon, Loevinger, Nouveauté, Intérêt, Dépendance causale, Leverage, Spécificité, Confirmation causale, Confiance causale, Confiance-confirmée causale, et Fiabilité négative*. Ces mesures appartiennent aux classes C_4, C_7 et G_{p8}, décrites dans le *chapitre 3*. Elles partagent les 8 propriétés suivantes : P_4, P_6, P_7, P_8, $\overline{P_{11}}$, $\overline{P_{14.2}}$, $\overline{P_{15}}$ et P_{21}.

Interprétation. Le facteur s'applique à un ensemble de mesures qui croissent en fonction du nombre d'exemples et en fonction de l'ensemble des données mais qui décroissent en fonction du nombre de motifs satisfaisant le conséquent. Il s'applique aussi aux mesures uniquement discriminantes, qui sont indifférentes aux premiers contre-exemples et possèdent un point variable dans le cas de l'équilibre.

Comparaison. La classification résultant de la méthode des *k-moyennes* (*présentée dans le chapitre 3, page 86*) révèle la présence de la classe G_{p8} qui regroupe les mesures *Piatetsky-Shapiro* et *VT100* de F_6. En revanche, selon la méthode de classification ascendante hiérarchique (*décrite dans le chapitre 3, page 85*), nous constatons que *VT100* est proche de *Précision* et à d'autres mesures complémentaires appartenant à la fois à C_4 et F_6, *e.g., Fiabilité négative, Cosinus* ou *Rappel*. Par ailleurs, nous signalons que les mesures de G_{p8} sont partagées entre F_1 (*Corrélation, Cohen, Force collective, Ratio des chances et Nouveauté*) et F_6 (*Piatetsky-Shapiro et VT100*).

Facteur 8. F_8 s'applique à 7 mesures, à savoir : *IPEE, IP3E, II, IPD, IIE, Indice de vraisemblance, IIER*. Toutes ces mesures partagent les 10 propriétés suivantes : P_4, P_5, P_6, $\overline{P_{10}}$, $\overline{P_{14.1}}$,

$\overline{P_{16}}$, $\overline{P_{917}}$, $\overline{P_{18}}$, P_{19}, P_{20}.

Interprétation. Le facteur s'applique à des mesures croissantes en fonction du nombre d'exemples mais qui n'attribuent pas de valeur fixe aux règles avec une confiance égale à 1. Il s'applique également à des mesures statistiques qui sont basées sur un certain modèle probabiliste.

Comparaison. F_8 correspond à deux classes décrites dans la classification consensuelle du chapitre 3, C_1 et C_2. Il contient 6 mesures et représente uniquement des mesures statistiques. Une seule mesure est manquante, *IIER*, qui selon la méthode de partitionnement (*k-moyennes*) appartient à C_1, et selon la méthode hiérarchique à C_2.

Facteur 10. La seule mesure restante qui n'a pas été couverte par les facteurs précédents, est *Fukuda*. Cette mesure (*Fukuda*) y compris les 15 mesures suivantes : *Confiance causale, Confiance-confirmée causale, Pavillon, Confirmation causale, Couverture, Dépendance, Dépendance causale, Loevinger, Indice d'implication,* M_{GK}*, Prévalence, Rappel, Fiabilité négative, Leverage* et *Spécificité* sont couvertes par le facteur F_{10}. Toutes ces mesures partagent les 7 propriétés suivantes : P_3, $\overline{P_{11}}$, $P_{14.2}$, $\overline{P_{15}}$, $\overline{P_{16}}$, $\overline{P_{18}}$ et $\overline{P_{19}}$.

Interprétation. Le facteur s'applique à des mesures non symétriques, non constantes au niveau de l'équilibre. Il s'applique aussi à des mesures qui ne sont pas basées sur un modèle probabiliste et qui sont indifférentes aux premiers contre-exemples.

Comparaison. Le facteur F_{10} couvre les classes C_3, C_4, C_6 et C_7, décrites dans le *chapitre 3* plus la mesure *Fukuda* pour laquelle aucun consensus n'a été trouvé. Cette mesure appartient à la classe C_3 selon la méthode hiérarchique (*CAH*), et à C_4 selon la méthode des *k-moyennes*. L'union des classes C_3, C_4, C_6 et C_7 dévoile *36* mesures d'intérêt parmi lesquelles une vingtaine sont absentes de F_{10}.

Pour seulement sept propriétés de mesures partagées, F_{10} représente le facteur ayant le nombre minimal de propriétés, communes à l'ensemble de ses mesures. C'est ce qui explique la dissemblance mutuelle des mesures appartenant à ce facteur.

Nous finalisons dans ce qui suit notre démarche par une discussion sur les résultats de la classification des mesures.

4.5 Discussion

L'étude que nous avons menée, dans ce chapitre, nous a permis d'avoir une classification différente de celle obtenue dans le *chapitre 3*. Une nouvelle structure des mesures est obtenue par l'application de l'AFB sur la matrice booléenne évaluant les mesures selon les propriétés. Cette structure, illustrée dans la *figure 4.2*, nous montre les mesures qui peuvent appartenir

à plusieurs classes à la fois, comme par exemple les mesures {*Intérêt, Risque relatif, Facteur bayésien, Support sens unique, Support double sens*} qui appartiennent à la fois à F_1 et F_5. Ces mesures partagent des propriétés communes avec les mesures des deux facteurs, ce qui a provoqué ce chevauchement entre les facteurs.

Toutefois, la discussion et la comparaison des résultats de la classification révélés par nos 2 expériences s'avère indispensable afin de discerner les groupes de mesures stables, *i.e,* les mesures qui quelque soit la méthode appliquée sont toujours groupées ensemble. Ainsi, nous reprenons le diagramme de Venn de la *figure 4.2*, qui représente les classes les plus importantes obtenues par l'AFB, et les 7 classes de mesures de la *figure 3.2*, obtenues suite à un consensus entre les méthodes de classification sans recouvrement, *CAH* et *k-moyennes*. La confrontation de ces deux structures, nous dévoile ces 8 groupes stables de mesures d'intérêt :

- G_{s1} = {*II, IVL*},
- G_{s2} = {*IIE, IPD, IP3E, IPEE*},
- G_{s3} = {*Gini, J-mesure, Prévalence, Information mutuelle, Indice d'implication, Dépendance, Couverture*},
- G_{s4} = {*Jaccard, Kulczynski, Czekanowski-dice, Spécificité, Leverage, Cosinus, Rappel, Dépendance causale*},
- G_{s5} = {*Czekanowski-dice, Spécificité, Leverage, Cosinus, Rappel, Dépendance causale, Confirmation causale, Confiance causale, Support, Confiance-confirmée causale, Fiabilité négative*},
- G_{s6} = {*Sebag, Taux d'exemples, Laplace, Confiance, Moindre contradiction, Ganascia, Confirmation descriptive*},
- G_{s7} = {*Y de Yule, Q de Yule, Zhang,* M_{GK}},
- G_{s8} = {*Intérêt, Gain informationnel, Risque relatif, Facteur bayésien, Conviction, Facteur de certitude, Pavillon, Klosgen, Support sens unique, Support double sens*}.

Nous discutons dans ce qui suit les 8 groupes stables de mesures $\{G_{s1}, \cdots, G_{s8}\}$.

Le premier groupe stable G_{s1} correspond à la classe C_1 de la *figure 3.2*, donc $G_{s1} = C_1$. De même pour G_{s2} qui est équivalent à la classe C_2. Nous notons que l'AFB réunit les deux groupes stables G_{s1} et G_{s2}, ou encore les deux classes C_1 et C_2 dans un même facteur, un rapprochement vérifiable à la consultation du dendogramme de la *figure 3.1*.

Le troisième groupe stable G_{s3} contient toutes les mesures de C_3, plus les mesures *Dépendance pondérée, Variation support* et *Pearl*, mesures appartenant à la classe C_3 selon l'une des méthodes de classification *CAH* ou *k-moyennes*. Ainsi, $G_{s3} = C_3 \bigcup$ {*Dépendance pondérée, Variation support et Pearl*}.

Les deux groupes stables G_{s4} et G_{s5} réunissent toutes les mesures de la classe C_4 de la manière suivante : $G_{s4} = C_4 -$ {*Confirmation causale, Confiance causale, Support, Confiance-confirmée causale, Fiabilité négative*}, et $G_{s5} = C_4 -$ {*Jaccard, Kulczynski*}.

Les groupes stables restants G_{s6}, G_{s7} et G_{s8} désignent respectivement les classes C_5, C_6 et C_7. Ainsi, nous avons $G_{s6} = C_5$, $G_{s7} = C_6$ et $G_{s8} = C_7$.

De ce point de vue, nous pouvons affirmer que l'AFB peut être considérée comme une méthode de classification complémentaire qui vient valider les résultats de la classification sans recouvrement. Les 8 groupes stables ainsi obtenus peuvent être utiles pour l'utilisateur qui désire sélectionner une (*ou des*) mesure(s), en particulier lorsque ce dernier a l'intention de choisir diverses mesures ayant des caractéristiques différentes, comme dans Bouker et al. [BSYN12]. Pour la forme particulière des classes utilisées dans la présente étude (*figure 4.2*), un tel scénario est particulièrement adapté parce que les groupes sont représentés par des concepts formels et donc possèdent une signification naturelle. En effet, la signification de chaque groupe provient de la collection d'attributs qui détermine uniquement la classe ou le concept formel correspondant(e).

Toutefois, les résultats obtenus par les deux travaux de classification réalisées sur les mesures d'intérêt ne se sont intéressés qu'au côté théorique. Ces études effectuées sur le comportement des mesures selon des propriétés formelles nécessitent d'être complétées par une étude selon un point de vue expérimental. De ce fait, une étude empirique s'avère indispensable pour analyser expérimentalement le comportement des mesures selon des jeux de données ainsi que pour vérifier si les mesures d'un même groupe stable de mesures, possèdent aussi empiriquement un comportement similaire.

4.6 Conclusion

Dans ce chapitre, nous avons analysé une soixantaine de mesures d'intérêt d'extraction de règles d'association en fonction de leurs propriétés binaires. Pour ce faire, nous avons utilisé une méthode récemment développée de l'analyse factorielle booléenne. Cette dernière a abouti à des facteurs booléens que nous avons interprétés comme étant des classes de mesures décrites par leurs propriétés. Nous avons vu également que ces facteurs booléens nous ont fournit des groupes de mesures qui sont clairement interprétables et significatifs.

Contrairement à d'autres méthodes de classification, telles que celles utilisées dans le *chapitre 3*, les facteurs booléens représentent des groupes non nécessairement disjoins. Ces groupes recouvrants ont l'avantage de former un phénomène naturel dans la classification de l'homme en général, et dans le cas où les objets sont représentés par des attributs empiétants en particulier.

Étant donnée la classification obtenue par l'AFB, nous avons effectué une comparaison détaillée de celle là avec la classification rapportée dans le chapitre précédent. Il s'est avéré que, parmi les groupes obtenus à partir des facteurs booléens, ceux correspondant au premier couple de facteurs sont très similaires à d'autres groupes de mesures discernés dans le *chapitre 3* par la classification sans recouvrement. D'autres études comparatives de ces classifications avec celles de travaux existants dans la littérature sont réalisées dans [BGG+13].

Les résultats de la classification obtenus dans le présent chapitre peuvent donc être considérés comme des résultats de validation des classes disjointes ainsi qu'ils peuvent être pris comme des points de référence pour de futures études sur la classification des mesures d'intérêt. Dans ce chapitre, nous avons également confronté les deux résultats de la classification obtenus (*comparer le diagramme de Venn avec les* 7 *classes de mesures du chapitre 3*) et nous avons détecté 8 groupes stables de mesures d'intérêt, i.e., des mesures qui sont toujours regroupées ensemble selon les 2 classifications.

Ainsi, l'étude menée sur les mesures a permis l'identification de groupes de mesures recouvrants d'une part et à des groupes stables d'autre part. Cependant, une étude empirique sur les mesures d'intérêt est nécessaire pour valider les résultats de la classification que nous avons obtenus : c'est l'objet du chapitre suivant.

Points clés

Positionnement :

– *Classification d'une soixantaine de mesures d'intérêt en utilisant des méthodes de classification avec recouvrement*.

Contribution :

– *Classification des mesures en utilisant une méthode de classification avec recouvrement : AFB* ;
– *Obtention de groupes de mesures qui se chevauchent, qui ont été confrontés aux* 7 *classes pour détecter* 8 *groupes de mesures stables*.

Publications :

– *R. Belohlavek, D. Grissa, S. Guillaume, E. Mephu Nguifo and J. Outrata (2011). Boolean factors as a means of clustering of interestingness measures of association rules*. CLA'2011, Nancy-FRANCE. pages 207–222.
– *R. Belohlavek, D. Grissa, S. Guillaume, E. Mephu Nguifo and J. Outrata (2013). Boolean factors as a means of clustering of interestingness measures of association rules*. AMAI Journal, volume 67, Springer Netherlands.

CHAPITRE 5

Étude empirique des mesures d'intérêt

Sommaire

5.1 Introduction

Les travaux de synthèse réalisés sur les différentes mesures objectives rencontrées dans la littérature, ont fait une comparaison de celles-ci selon plusieurs points de vue : (i) étude formelle

suite à la définition d'un ensemble de propriétés de mesures, qui conduisent à une bonne évaluation de celles-ci [HH01], [TKS02], [LT04], [GH07], [Fen07] ; (ii) étude comparative expérimentale du comportement des différentes mesures d'intérêt à partir du point de vue d'analyse de données [VLL04], [HGB05a], [CFE05], [HZ10], [SKR10]. Dans les chapitres précédents, nous avons effectué une étude formelle sur 61 mesures d'intérêt décrites par 19 propriétés, où des classes de mesures ont été identifiées. Néanmoins, envisager une approche complémentaire par une étude expérimentale de ces mesures nous semble important, puisqu'elle permettra d'analyser empiriquement le comportement de ces mesures et de découvrir celles qui se ressemblent. Cette approche est similaire à celle effectuée par [Vai06], mais diffère au niveau du nombre de mesures et des jeux de données étudiés (20 *mesures contre* 61 *mesures*).

Dans le présent chapitre, nous menons une étude expérimentale sur le comportement des différentes mesures d'intérêt. Cette étude consiste à analyser les *N* meilleures règles extraites dans des bases de données différentes, et par chacune des mesures, afin de s'assurer que cet ensemble des *N* meilleures règles est sensiblement le même dans chacune des classes extraites par l'étude formelle menée dans le *chapitre 3*. Le choix des bases est aussi une étape très importante dans cette étude puisque nous sommes bien conscients que la catégorisation des mesures dépend nécessairement de plusieurs facteurs (*e.g., les données ou l'expert-utilisateur*), comme le souligne [Suz08]. Pour éviter le biais des données, nous choisissons de varier les jeux de données et nous appliquons nos expérimentations sur deux types de bases distincts : un jeu de données "*réelles*" et un jeu de données "*synthétiques*".

Ce chapitre est organisé de la manière suivante : nous entamons par une description de la méthodologie expérimentale. Nous analysons par la suite les caractéristiques des jeux de données. Puis, nous exposons les résultats de l'étude expérimentale réalisée sur les mesures d'intérêt par le biais de bases de nature différente. Dans la section suivante, nous cherchons à interpréter les résultats empiriques obtenus avant de les confronter ensuite avec les résultats de l'approche formelle. Enfin, nous proposons de vérifier la classification retenue par l'étude expérimentale avec celles obtenues par [Vai06], [HGB+07] et [Bra11].

5.2 Méthodologie expérimentale

Dans cette section, nous cherchons à étudier empiriquement le comportement des mesures d'intérêt d'extraction de règles d'association. Pour ce faire, nous proposons une démarche expérimentale que nous illustrons tout d'abord dans le diagramme de la *figure 5.1* pour présenter les différentes étapes suivies. Les notations incluses dans ce diagramme sont présentées en détail dans le descriptif de la méthodologie ci-dessous.

Nous expliquons et décrivons dans ce qui suit le processus de la *figure 5.1*, composé de 6

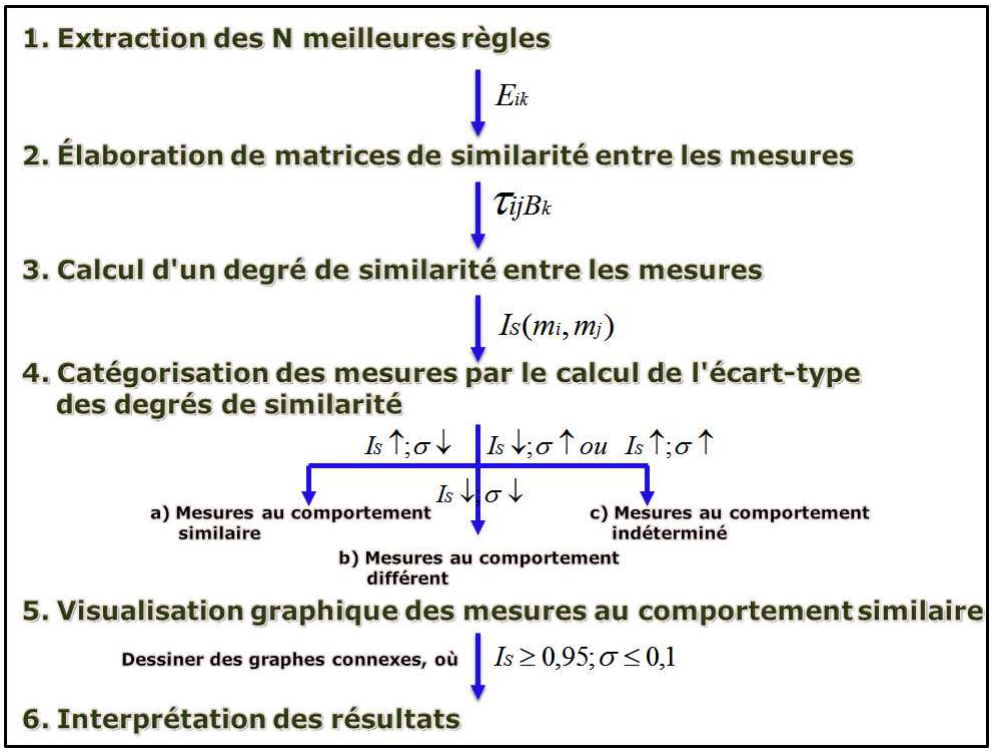

FIGURE *5.1: Diagramme de la méthodologie expérimentale.*

étapes :

- **Étape 1 : Extraction des *N* meilleures règles**

  Nous appliquons pour chaque ensemble de données et pour chaque mesure d'intérêt étudiée, l'algorithme "*Apriori*". L'utilisation de cet algorithme n'est qu'un paramètre, il est possible d'utiliser d'autres outils, voir [WKQ+07].

  À l'issue de cette extraction, nous obtenons pour chacune des mesures, un ensemble des *N* meilleures règles E_{ik} jugées pertinentes par la mesure m_i avec $(i = 1, \cdots, c)$ pour la base B_k $(k = 1, \cdots, p)$, sachant que c représente le nombre de mesures d'intérêt $(c = 61)$ et p désigne le nombre de bases analysées $(p = 6)$. Les *N* meilleures règles sélectionnées représentent les règles les mieux classées par une mesure m_i, i.e., la mesure m_i leur attribue les valeurs les plus élevées.

  Nous obtenons à l'issue de cette première étape, un tableau décrivant l'ensemble E_{ik} des règles extraites par la mesure m_i pour la base B_k.

- **Étape 2 : Élaboration de matrices de similarité entre les mesures**

  Pour chaque base de données B_k $(k = 1, \cdots, p)$, nous comparons les ensembles E_{ik} et E_{jk} des *N* meilleures règles extraites par les mesures d'intérêt m_i et m_j. Au cours de cette comparaison, nous nous intéressons uniquement à la présence des règles dans les deux ensembles et non pas à l'ordre avec lequel les mesures classent ces règles. Nous

jugeons que deux mesures se ressemblent si elles extraient les mêmes *N* meilleures règles, quelque soit l'ordre dans lequel elles se présentent. La présence étant plus significative que l'ordre pour juger la similarité de deux mesures.

La comparaison des deux ensembles de règles est alors effectuée afin de calculer le pourcentage de règles communes à ces deux ensembles permettant ainsi d'obtenir un taux de ressemblance entre les deux mesures. Le taux de ressemblance $\tau_{ij\,B_k}$ des mesures m_i et m_j pour la base B_k se calcule de la façon suivante :

$$\tau_{ij\,B_k} = \frac{\mid E_{ik} \cap E_{jk} \mid}{N} \tag{5.1}$$

où $\mid E_{ik} \cap E_{jk} \mid$ est la cardinalité de $E_{ik} \cap E_{jk}$, c'est-à-dire l'ensemble de règles extraites par à la fois la mesure m_i et la mesure m_j.

Nous obtenons à l'issue de cette deuxième étape, un ensemble de p matrices M_k de taux de ressemblance entre chaque paire de mesures.

– **Étape 3 : Calcul d'un degré de similarité entre les mesures**

À partir des p matrices de taux de similarité obtenues à l'étape précédente, nous allons calculer un degré de similarité I_S pour chaque couple de mesures ($m_i,\ m_j$) de la façon suivante :

$$I_S(m_i, m_j) = \frac{\sum_{k=1}^{p} \tau_{ij\,B_k}}{p} \tag{5.2}$$

Une nouvelle matrice de similarité est ainsi obtenue décrivant la moyenne des p taux de ressemblance $\tau_{ij\,B_k}$ entre couple de mesures pour tous les jeux de données étudiés.

Puisque les mesures dépendent de la nature des données, il nous semble important de prendre ce facteur en considération dans notre étude et d'étudier l'écart-type des taux de ressemblance τ.

– **Étape 4 : Catégorisation des mesures par le calcul de l'écart-type des taux de ressemblance**

Pour chaque couple (m_i, m_j) de mesures, nous calculons l'écart-type $\sigma(m_i, m_j)$ des taux de ressemblance τ_{ij} afin de détecter les couples de mesures stables, i.e., qui possèdent un comportement similaire par la proposition des mêmes *N* meilleures règles et ceux qui ne le sont pas. Ces mesures stables sont identifiées pour une faible valeur de l'écart-type et une forte valeur du degré de similarité (*proche de 1*). Ce calcul s'effectue de la façon suivante :

$$\sigma(m_i, m_j) = \sqrt{\frac{\sum_{k=1}^{p} \left[\tau_{ij\,B_k} - I_S(m_i, m_j)\right]^2}{p}} \tag{5.3}$$

À l'issue de cette étape, nous cherchons à catégoriser les différentes mesures grâce au degré de similarité I_S et l'écart-type σ. Ainsi, on peut dégager les 3 catégories de mesures suivantes :

1. *mesures au comportement similaire :* lorsque le degré de similarité est proche de 1 et l'écart-type est faible ;
2. *mesures au comportement différent :* lorsque I_S et σ ont des faibles valeurs ;
3. *mesures au comportement indéterminé, en fonction des bases de données :* lorsque (i) I_S a une valeur forte et σ une valeur faible, (ii) I_S et σ ont tous deux une valeur forte.

La formalisation de ces différentes catégories est présentée dans la *table 5.1*.

Catégorie	**Indice** I_S	**Écart-type** σ
mesures au comportement similaire	$I_S(m_i, m_j) \geq 1 - \varepsilon_1$	$\sigma(m_i, m_j) \leq \varepsilon_2$
mesures au comportement différent	$I_S(m_i, m_j) \leq \varepsilon_1$	$\sigma(m_i, m_j) \leq \varepsilon_2$
mesures au comportement indéterminé	—	$\sigma(m_i, m_j) \geq 1 - \varepsilon_2$

TABLE *5.1: Étude du comportement des mesures par le calcul du degré de similarité et de l'écart-type.*

– **Étape 5 : Visualisation graphique des mesures au comportement similaire**
Cette étape s'intéresse à la visualisation des mesures au comportement similaire, qui appartiennent à la *catégorie 1*. Pour ce faire, nous proposons de dessiner des graphes connexes [1] G_l pour chaque groupe de mesures identifié. Ainsi, nous nous appuyons sur les deux matrices obtenues dans les *étapes 3 et 4* afin de chercher les valeurs de I_S et σ supérieures à deux seuils ε_{I_S} et ε_σ fixés. Ces deux seuils représentent respectivement $1 - \varepsilon_1$ et ε_2 et valent $0,95$ et $0,1$ (*le choix de ces valeurs est expliqué dans la sous-section 5.4.3, page 156*). Ainsi, nous jugeons proches les mesures ayant un degré de similarité $I_S \geq 0,95$ et un écart-type $\sigma \leq 0,1$. Les groupes de mesures qui vérifient ces deux contraintes sont appelés stables, i.e., qui possèdent un comportement semblable. Ces groupes sont visualisables par des graphes complets dont les arêtes sont étiquetées par I_S et σ.
– **Étape 6 : Interprétation des résultats**
Cette dernière étape de notre méthodologie expérimentale consiste à interpréter les résultats de la catégorisation des mesures obtenus précédemment. Nous nous focalisons essentiellement sur les groupes de mesures stables (*catégorie 1*), qui possèdent un

1. Le graphe G est dit connexe lorsqu'il existe une chaîne entre deux sommets quelconques de G.

comportement semblable quelque soit le jeu de données sélectionné.

Avant de suivre cette démarche expérimentale, nous présentons les différentes bases de données qui vont servir à déterminer les diverses catégories de mesures.

5.3 Étude des jeux de données

Nous utilisons 6 bases de données de 2 types différents : 4 bases de données réelles et 2 bases de données synthétiques. En outre, pour augmenter la diversité des situations possibles, nous avons sélectionné également des bases denses [2] [GZ01] et éparses [3].

Les bases de données réelles retenues sont les suivantes : *CHESS*, *CONNECT* et *PUMSB* qui proviennent du répertoire FIMI (*Frequent Itemset Mining Implementations*) [4], et *IPUMS* qui est issue de l'archive d'UCI KDD [HB99].

Les bases de données synthétiques sont les suivantes : *T135L23I60* et *T100L10I40*. Elles sont produites par le générateur d'*IBM* [5] à partir de différents paramètres qui sont les suivants : $|T|$ qui est le nombre de transactions, $|L|$ correspondant à la taille moyenne des transactions et $|I|$ symbolisant le nombre d'items. Ainsi, la base *T135L23I60* est composée de 135 transactions (*T=135*), décrites par 60 items (*I=60*) avec une moyenne de 23 items par transaction (*L=23*).

La *table 5.2* résume les caractéristiques des jeux de données utilisés pour cette étude. Ce tableau définit pour chaque base : (1) son type (*D : Dense, é : éparse, ? : ni dense ni éparse, R : Réel et S : Synthétique*), (2) T : le nombre de transactions, (3) I : le nombre d'items et (4) L : le nombre moyen d'items par transaction.

Nom de la base	Type	T :Nb.Tran	I :Nb.Item	L :Moy.Tran
CONNECT	D/R	67557	129	74
PUMSB	D/R	49046	7117	43
CHESS	D/R	3196	75	37
IPUMS	?/R	88443	1889	60
T135L23I60	é/S	135	60	23
T100L10I40	é/S	100	40	10

TABLE *5.2: Caractéristiques des 6 jeux de données.*

2. Une base de données est dite "*dense*" si elle produit des motifs fréquents de grande taille même pour des valeurs élevées du seuil minimum du support.

3. Une base de données est dite "*éparse*" si ses données sont faiblement corrélées, telles que les données synthétiques.

4. http ://fimi.ua.ac.be/data/

5. http ://ibmquestdatagen.sourceforge.net/

Les différentes expérimentations sont réalisées au moyen de l'environnement de travail WEKA [WF00], un support utile pour nous aider à l'analyse des mesures. Afin de mener à bien nos expérimentations, des paramètres importants ont été introduits à travers l'interface graphique de cet outil, qui sont les suivants :

- *delta* : ce facteur fait décroître le support minimal, jusqu'à ce que soit le nombre de règles demandées est trouvé, soit on a atteint la valeur minimale du support ;
- *lowerBoundMinSupport* : qui désigne la valeur minimale du support min_{sup}. Cette valeur dépend du jeu de données sélectionné ;
- *metricType* : ce paramètre désigne la mesure d'intérêt qui permet de classer les règles, tels que la *confiance*, *Leverage*, etc. À ce niveau, nous avons introduit une cinquantaine de mesures dans la plateforme Weka ;
- *minMetric* : c'est la valeur minimale de la mesure en dessous de laquelle on ne recherchera plus de règle. Elle diffère d'une mesure à une autre ;
- *numRules* : désigne le nombre de règles que l'algorithme doit produire. Nous avons travaillé avec un nombre *N = 100* mais nous l'avons aussi varié (*pour 2 jeux de données*) afin d'étudier son effet sur les résultats ;
- *upperBoundMinSupport* : c'est la valeur initiale du support qui décroît conformément à delta.

Afin d'étudier l'effet de la variation du support sur le comportement des mesures, nous choisissons différentes valeurs pour le paramètre *lowerBoundMinSupport*, et ce pour les deux bases de données, IPUMS et CHESS, uniquement. La *table 5.3* décrit un extrait des valeurs du taux de ressemblance τ_{ij} entre certains couples de mesures, prises pour le cas de la base CHESS en variant les valeurs du support minimum. Cependant, pour d'obtenir des résultats avec des temps d'extraction raisonnables, nous retenons un support minimum suffisant élevé. Les résultats ainsi obtenus montrent que les mesures se comportent toujours de la même façon quelque soit la valeur retenue pour le support minimum, montrant ainsi l'absence du biais du support.

S'appuyant sur ces jeux de données, nous exposons dans ce qui suit les résultats de la série d'expérimentations réalisées sur une soixante de mesures d'intérêt.

5.4 Résultats expérimentaux

Nous entamons nos expérimentations par l'application des deux premières étapes (*étape 1 et étape 2*), présentées dans la *section 5.2*, sur les deux types de bases : les bases réelles et les

min_sup	65%	75%	80%	85%
Lap&Conf	0,98	0,98	0,98	0,98
Lap&Pearl	0,00	0,00	0,00	0,00
Lap&MoCo	0,32	0,32	0,32	0,32
Gan&Conf	0,98	0,98	0,98	0,98
Gan&Pearl	0,00	0,00	0,00	0,00
Gan&MoCo	0,32	0,32	0,32	0,32
Conf&Pearl	0,00	0,00	0,00	0,00
Conf&MoCo	0,32	0,32	0,32	0,32
Pearl&MoCo	0,00	0,00	0,00	0,00

TABLE *5.3: Valeurs du taux de ressemblance entre couples de mesures en variant le support minimum de la base CHESS.*

bases synthétiques. Pour ce faire, nous fixons d'abord le nombre de règles à extraire par chaque mesure, et nous avons choisi *N = 100*, un nombre de règles qui semble raisonnable pouvant être exploré par un utilisateur [CN06]. Nous aurions pu choisir un nombre plus élevé pour la variable *N*, mais lorsque nous obtenons les 100 premières règles quasiment identiques pour 2 mesures différentes, il en est bien souvent de même pour les règles suivantes puisque cette similarité nous indique une évaluation semblable des règles extraites par ces deux mesures, comme le montre la *table 5.4*.

Afin d'étudier l'effet de la variation de *N* sur les résultats, nous présentons dans la *table 5.4*, à partir d'un sous-ensemble des *N* meilleures règles (*où N varie entre* 10 *et* 400), le nombre de règles communes entre 10 couples de mesures dans le cas de la base CHESS. La taille de ces sous-ensembles est au moins égale à *N* pour les mesures qui ordonnent les règles de la même manière.

N	**10**	**50**	**100**	**200**	**400**
Lap&Gan	10	50	100	200	400
Lap&Conf	10	50	98	184	366
Lap&Pearl	0	0	0	0	0
Lap&MoCo	7	30	32	32	36
Gan&Conf	10	50	98	184	366
Gan&Pearl	0	0	0	0	0
Gan&MoCo	7	30	32	32	36
Conf&Pearl	0	0	0	0	0
Conf&MoCo	7	30	32	32	35
Pearl&MoCo	0	0	0	0	0

TABLE *5.4: Nombre des meilleures règles communes à partir du sous-ensemble des N meilleures règles de la base CHESS.*

Nous remarquons à partir de cette table, que la variation du nombre *N* de règles que l'utilisateur peut examiner provoque des différences pas vraiment importantes dans les résultats,

d'où une faible variation de la similarité.

Par la suite et à l'issue de cette extraction des 100 meilleures règles (*étape 1*) sur nos 6 bases, nous comparons les 6 ensembles E_{ik} et les 6 ensembles E_{jk} ($k = 1, \cdots, 6$), et ce pour tous les couples de mesures (m_i, m_j) afin d'obtenir 6 matrices M_k de similarité (*étape 2*) indiquant le taux de ressemblance de ces mesures. Nous illustrons notre démarche à travers deux petits exemples, pour chaque type de base de données (*réelle ou synthétique*), mettant en jeu 5 mesures d'intérêt. Nous commençons par le cas des bases réelles.

5.4.1 Cas des ensembles de données réelles

À partir des 4 bases de données réelles étudiées, nous prenons comme exemple, le cas de la base dense "CONNECT". Nous illustrons dans la *table 5.5*, les valeurs du taux de ressemblance $\tau_{ij\,B_k}$ entre les 5 mesures d'intérêt suivantes : *Conv : Conviction, FB : Facteur bayésien, FC : Facteur de certitude, FN : Fiabilité négative, CC : Confiance causale*. Ainsi, nous découvrons qu'il existe des couples de mesures ayant un taux de ressemblance $\tau_{ij} = 1$, tel que le couple {*Facteur bayésien, Conviction*} ; ce qui signifie que ces deux mesures possèdent un comportement identique suite à l'extraction des mêmes 100 meilleures règles E_{ik}. D'autres couples de mesures, comme {*Fiabilité négative, Facteur de certitude*} possèdent un comportement très différent puisque la valeur du taux de ressemblance est soit nulle ($\tau_{ij} = 0$), soit faible. Ces mesures possèdent ainsi deux ensembles de règles E_{ik} et E_{jk} dissemblables.

Mesure	Conv	FB	FC	FN	CC
Conv	$1,00$				
FB	**1,00**	$1,00$			
FC	$0,00$	$0,00$	$1,00$		
FN	$0,91$	$0,91$	**0,00**	$1,00$	
CC	$0,91$	$0,91$	$0,00$	$1,00$	$1,00$

TABLE *5.5: Extrait de la matrice de similarité entre* 6 *couples de mesures pour la base "Connect".*

Au vu des résultats obtenus par l'étude expérimentale réalisée sur les 4 bases réelles, nous remarquons qu'il existe différentes catégories de mesures, e.g., des mesures qui ont un comportement très semblable et d'autres qui varient de comportement selon la base de données sélectionnée.

5.4.2 Cas des ensembles de données synthétiques

Pour le cas des bases synthétiques, nous procédons de la même manière que précédemment, en suivant les deux premières étapes de notre méthodologie expérimentale. Une série de

tests est ainsi réalisée sur les 2 bases synthétiques "*T100L10I20*" et "*T135L23I60*" permettant d'obtenir 2 différentes matrices de similarité des mesures M_k. Nous retenons dans la *table 5.6* un extrait de la matrice correspondante à la base "*T100L10I20*" pour les mêmes 5 mesures d'intérêt retenues dans la section précédente. En comparant ces mesures deux à deux, nous remarquons comme pour le cas des bases de données réelles, la présence de mesures ayant un même comportement, e.g., {*Confiance causale, Fiabilité négative*} ou {*Facteur bayésien, Conviction*} pour un taux de ressemblance $\tau_{ij} = 1$ ou proche de 1 $(0,91)$; et de mesures qui ont un comportement différent, avec une faible valeur du taux de ressemblance égale à 0 ou proche de 0, tel est le cas des couples de mesures {*Facteur de certitude, Facteur bayésien*} ou {*Fiabilité négative, Facteur de certitude*}.

Mesure	Conv	FB	FC	FN	CC
Conv	1.00				
FB	**0.91**	1.00			
FC	0.02	**0.00**	1.00		
FN	1.00	0.91	**0.02**	1.00	
CC	1.00	0.91	0.02	**1.00**	1.00

TABLE *5.6: Extrait de la matrice de similarité entre 6 couples de mesures pour la base "T100L10I20".*

Les résultats révélés par les deux matrices de similarité des 2 bases synthétiques montrent la présence de mesures qui se comportent de la même façon, d'autres qui dépendent du jeu de données sélectionné et de mesures qui possèdent un comportement très différent.

À partir des résultats obtenus pour les deux types de données, nous concluons qu'il n'est pas possible de catégoriser les mesures selon la nature des données ou de la densité de la base, mais plutôt en fonction d'autres paramètres, comme le montre la section suivante.

5.4.3 Catégorisation des mesures

Nous continuons notre travail par l'application des deux étapes suivantes : *étape 3* et *étape 4* (*présentées dans la section 5.2*), que nous illustrons à travers un petit exemple mettant en jeu les 5 mesures d'intérêt, citées dans les deux sections précédentes. Ainsi, la *table 5.7* restitue les valeurs du degré de similarité I_S pour les couples de mesures suivants : {*Conviction, Facteur bayésien*}, {*Conviction, Facteur de certitude*}, {*Conviction, Confiance causale*}, {*Fiabilité négative, Facteur bayésien*}, {*Fiabilité négative, Facteur de certitude*} et {*Fiabilité négative, Confiance causale*} ainsi que leur écart-type σ.

Nous fixons respectivement pour ε_1, ε_2 les valeurs suivantes : $0,05$ et $0,1$. Le choix de ces seuils n'est pas fait au hasard mais après vérification des valeurs de I_S et σ entre couples

de mesures, à partir des deux matrices obtenues dans les *étapes 3 et 4*. Nous tolérons de baisser les valeurs des seuils ε_1 et ε_2 afin de considérer les couples de mesures tels que {*Facteur bayésien, Conviction*}, ayant comme valeurs pour $I_S = 0,95$ et $\sigma = 0,06$. De telles mesures sont jugées similaires puisqu'elles extraient 100 règles identiques pour les bases "IPUMS, CONNECT et T135L23I60" et 91 règles pour la base "CHESS". Par conséquent, nous jugeons les seuils fixés de ε_1, ε_2, comme étant de bons estimateurs pour identifier les catégories de mesures. D'après les critères retenus à l'*étape 4* pour constituer les 3 catégories de mesures, nous pouvons en déduire les similarités, dissimilarités ou indéterminations suivantes pour ces 6 mesures :

- *Catégorie 1* : mesures au comportement similaire, c'est le cas des 2 couples de mesures {*Conviction, Facteur bayésien*} et {*Fiabilité négative, Confiance causale*} ;
- *Catégorie 2* : mesures au comportement différent, tels que les couples {*Conviction, Facteur de certitude*} et {*Fiabilité négative, Facteur de certitude*} ;
- *Catégorie 3* : indétermination, c'est le cas des couples de mesures {*Fiabilité négative, Facteur bayésien*} et {*Conviction, Confiance causale*}.

Mesure	**Conviction**		**Fiabilité Négative**	
-	I_S	σ	I_S	σ
Facteur Bayésien	$0,95$	$0,06$	$0,60$	$0,38$
Facteur de certitude	$0,00$	$0,00$	$0,00$	$0,00$
Confiance Causale	$0,58$	$0,37$	$1,00$	$0,00$

TABLE *5.7: Extrait des matrices représentant respectivement le degré de similarité entre* 6 *couples de mesures ainsi que l'écart-type.*

Remarque : Dans le *chapitre 2, section 2.5, page 71*, nous avons identifié des relations mathématiques entre les mesures. Certaines mesures ont montré des liens forts entre elles, comme nous l'avons expliqué dans la page 71. Ces mesures (*e.g., Intérêt, Gain informationnel*) peuvent être ignorées dans cette étude empirique du comportement des mesures, réduisant par conséquent la quantité des mesures analysées. Toutefois et afin de s'assurer de la similarité de leur résultats, nous tenons à les étudier expérimentalement.

Dans ce qui suit, nous nous intéressons à l'interprétation des groupes de mesures appartenant à la première catégorie de mesures.

5.5 Interprétation des groupes de mesures stables

Dans cette section, nous cherchons à identifier les groupes de mesures stables, i.e., qui extraient quasiment les mêmes meilleures règles quelque soit le jeu de données sélectionné, et

à les interpréter.

5.5.1 Visualisation des groupes de mesures stables

Comme nous l'avons déjà expliqué dans l'*étape 5* de notre méthodologie expérimentale, pour découvrir les groupes de mesures stables, nous avons recours à la fixation de deux seuils $\varepsilon_{I_S} = 1 - \varepsilon_1$ et $\varepsilon_\sigma = \varepsilon_2$. Comme $\varepsilon_1 = 0,05$ et $\varepsilon_2 = 0,1$, alors $\varepsilon_{I_S} = 0,95$ et $\varepsilon_\sigma = 0,1$. Une fois les seuils sont déterminés, nous reprenons les deux matrices obtenues dans les *étapes 3 et 4*, pour découvrir les groupes de mesures dont les valeurs de I_S et σ sont supérieures ou égales à ces deux seuils. Ainsi, nous identifions les 8 groupes de mesures stables suivants :

- G_{St1}= {*Pea (Pearl), Piatetsky-shapiro (PS), Nouveauté (Nov), Leverage (Lev)*} ;
- G_{St2}= {M_{GK}, *Facteur de certitude (FC), Taux d'exemples (TEC), Risque relatif (RR), Sebag (Seb), Laplace (Lap), Ganascia (Gan)*} ;
- G_{St3}= {*Information mutuelle (IM), Variation support (VS)*} ;
- G_{St4}= {*Gain informationnel (GI), Intérêt (Int)*} ;
- G_{St5}= {*Jaccard (Jac), Kulczynski (Kulz), Czekanowski-dice (CzD)*} ;
- G_{St6}= {*Confiance-confirmée causale (CCC), Confiance causale (Cconf), Fiabilité négative (FN)*} ;
- G_{St7}= {*Support (Sup), IIE, IIER, IP3E, IPEE, Rappel (Rap)*} ;
- G_{St8}= {*Facteur Bayésien (FB), Conviction (Conv)*}.

Afin de visualiser ces groupes de mesures stables, nous traçons dans la *figure 5.2* des graphes complets G_l avec ($l = 1, \cdots, 8$), pour chaque groupe discerné. Les noeuds de ces graphes représentent les mesures d'intérêt et les artères ne relient deux noeuds que lorsque le couple de mesures concerné vérifie les contraintes des valeurs du degré de similarité et de l'écart-type. Nous mentionnons également sur les arêtes (*ou arcs*) des graphes les valeurs de I_S et σ.

Nous procédons dans ce qui suit à l'interprétation de ces 8 groupes stables.

5.5.2 Les groupes de mesures stables

Afin de comprendre et d'interpréter ces différents groupes de mesures, nous allons suivre la démarche illustrée dans la *figure 5.3*, et que nous décrivons dans ce qui suit.

Étape a : *Recherche de relations mathématiques entre les mesures*. Nous allons rechercher des relations entre les mesures au moyen de la *table 2.5, page 73*, comme par exemple des liens de proportionnalité, afin de justifier ce regroupement. Dans le cas où nous ne trouvons pas

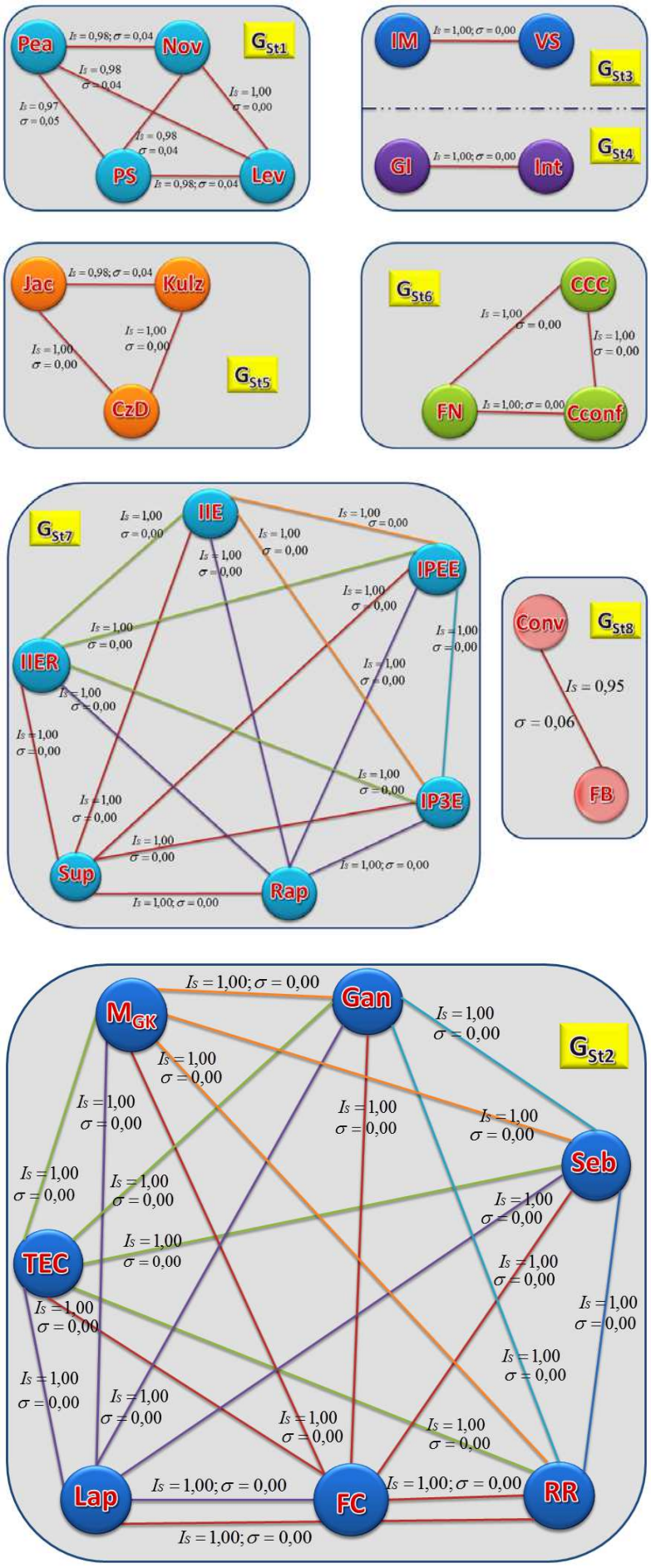

FIGURE *5.2: Les groupes de mesures stables obtenus par l'étude empirique.*

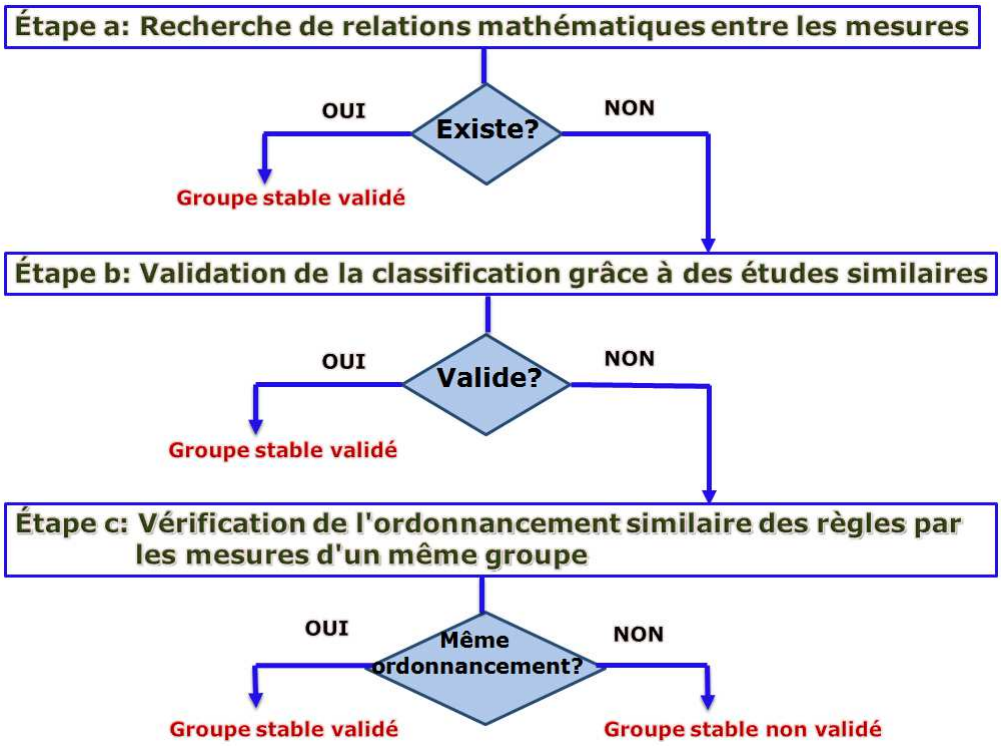

FIGURE *5.3: Diagramme illustrant la démarche méthodologique suivie pour l'interprétation des groupes de mesures.*

de relations intéressantes, nous passons à l'*étape b*, où nous allons comparer les groupes de mesures avec ceux obtenus lors de travaux similaires.

Étape b : *Validation de la classification grâce à des études similaires*. Si certains regroupements de mesures ou même la totalité n'ont pas pu être expliqués grâce à l'*étape a*, nous confrontons cette classification à celles obtenues dans la littérature (*de façon formelle ou expérimentale*) ou dans les *chapitres 3* et *4*. Dans le cas où de nouveau, ces résultats ne peuvent être validés, i.e., il n'est pas possible de vérifier la classification avec des travaux existants, nous passons à la dernière étape qui va consister à vérifier que ces mesures ordonnent de la même façon les règles.

Étape c : *Vérification de l'ordonnancement similaire des règles par les mesures d'un même groupe*. Nous allons vérifier que les mesures (m_1, m_2) d'un même groupe présentent des ensembles E_{1k} et E_{2k} identiques de règles ordonnées de la même manière. Ainsi, la relation des pré-ordres induits par les mesures suivante doit être vérifiée par le couple de mesures afin de s'assurer de leur comportement similaire :

$\forall X_1 \rightarrow Y_1,\ \forall X_2 \rightarrow Y_2 \quad Si \quad m_1(X_1 \rightarrow Y_1) \leq m_1(X_2 \rightarrow Y_2) \quad alors \quad m_2(X_1 \rightarrow Y_1) \leq m_2(X_2 \rightarrow Y_2)$

Pour cela, nous allons reproduire le cas des bases de données denses et éparses en pas-

sant par une situation intermédiaire ; et pour chacune des 3 situations précédentes, nous évaluons toutes les configurations possibles de règles à savoir l'état d'incompatibilité jusqu'à l'implication logique en passant notamment par l'indépendance et l'équilibre. Pour y parvenir, nous allons tracer les courbes d'évolution des mesures (*voir figure 1.4, page 25*) pour les 3 situations répertoriées dans les *figures 5.4 et 5.5*.

Pour une règle quelconque $X \rightarrow Y$, la *figure 1.4, page 25*, répertorie tous les états que l'on peut rencontrer, à savoir de l'état d'incompatibilité jusqu'à l'implication logique. Cette évolution du nombre d'individus vérifiant à la fois X et Y sera l'axe des abscisses de nos futures courbes (*e.g., figure 5.6*). Cette évolution de la *figure 1.4*, va être reproduite dans les 3 situations des *figures 5.4 et 5.5*, symbolisant les différents types de base de données que l'on peut rencontrer en variant respectivement le nombre d'individus total n et le nombre du conséquent n_Y.

Ainsi, pour une règle quelconque $X \rightarrow Y$, nous allons dans un premier temps simuler le cas de la base de données dense en choisissant une valeur pour le nombre d'individus total (*donc taille de la base de données*) assez proche des cardinalités des ensembles T_X et T_Y ; c'est le cas du diagramme de Venn de gauche de la *figure 5.4*. Le cas de la base de données éparse sera simulé au contraire par une valeur élevée pour $|T|$, représenté par le diagramme de Venn de droite de la *figure 5.4*. Par la suite, nous allons resimuler ces bases de données mais cette fois en variant le nombre du conséquent n_Y ou l'ensemble T_Y. Pour le cas de la base de données dense, nous choisissons une valeur pour n_Y assez proche de la cardinalité de l'ensemble T_X (*représentée par le diagramme de Venn de gauche de la figure 5.5*) et pour le cas de la base de données éparse (*représentée par le digramme de Venn de droite de la figure 5.5*), nous la simulons avec une valeur élevée pour $|T_Y|$ ou n_Y.

Remarque : les cardinalités des ensembles T_X, T_Y et T_{XY} sont toutes identiques pour les 3 diagrammes de la *figure 5.4*.

S'appuyant sur la démarche ainsi décrite, nous allons interpréter les 8 groupes stables de mesures retenus.

Groupe G_{St1} :

- *Étape a* : Selon la *table 2.5, page 73*, il existe une relation de proportionnalité entre le couple de mesures {*Nouveauté, Piatetsky-shapiro*} de la forme $Piatetsky - Shapiro = T \times |Nouveauté|$, avec T le nombre de transactions. Il s'agit de passer d'une mesure à l'autre en multipliant par une constante T non nulle. Cette relation montre que ces deux mesures possèdent un comportement similaire.
- *Étape b* : Seul le couple de mesures {*Nouveauté, Piatetsky-shapiro*} déjà vérifié dans l'étape précédente est validé dans cette étape selon la classification obtenue dans le *chapitre 3* (*cf. section 3.3, page 83*). Il est donc indispensable de passer à l'étape suivante

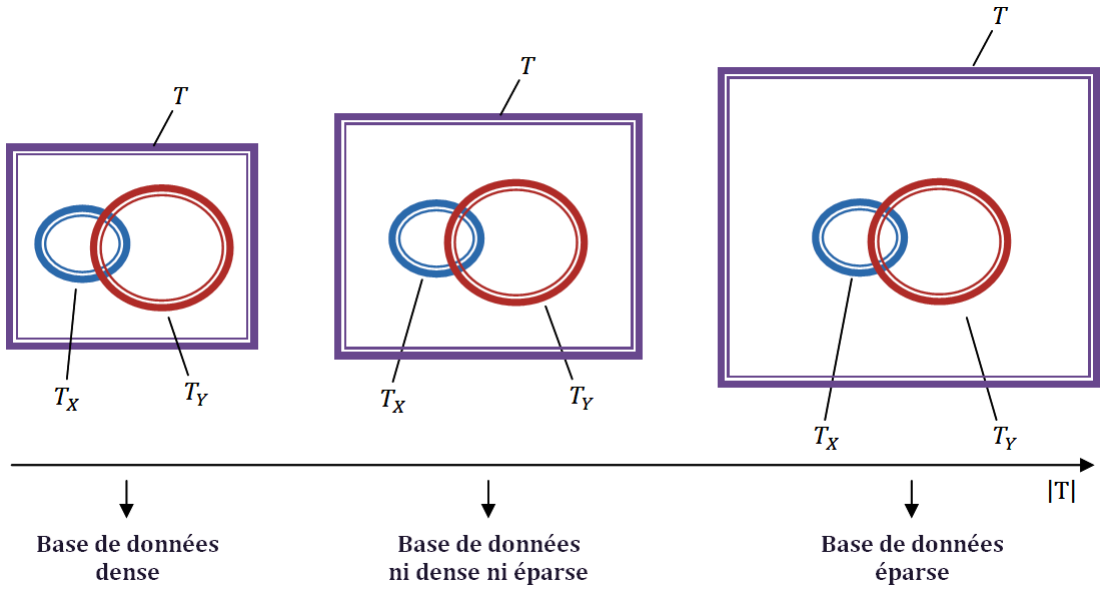

FIGURE *5.4: Les trois situations retenues simulant différents types de bases de données en variant le nombre total d'individus.*

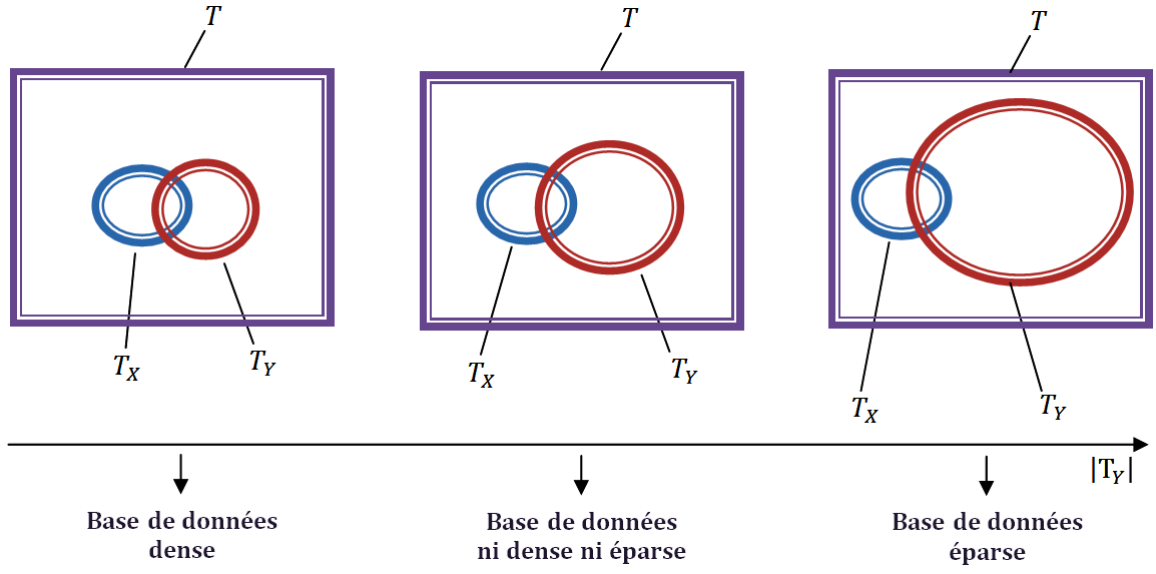

FIGURE *5.5: Les trois situations retenues simulant différents types de bases de données en variant la taille du conséquent* Y.

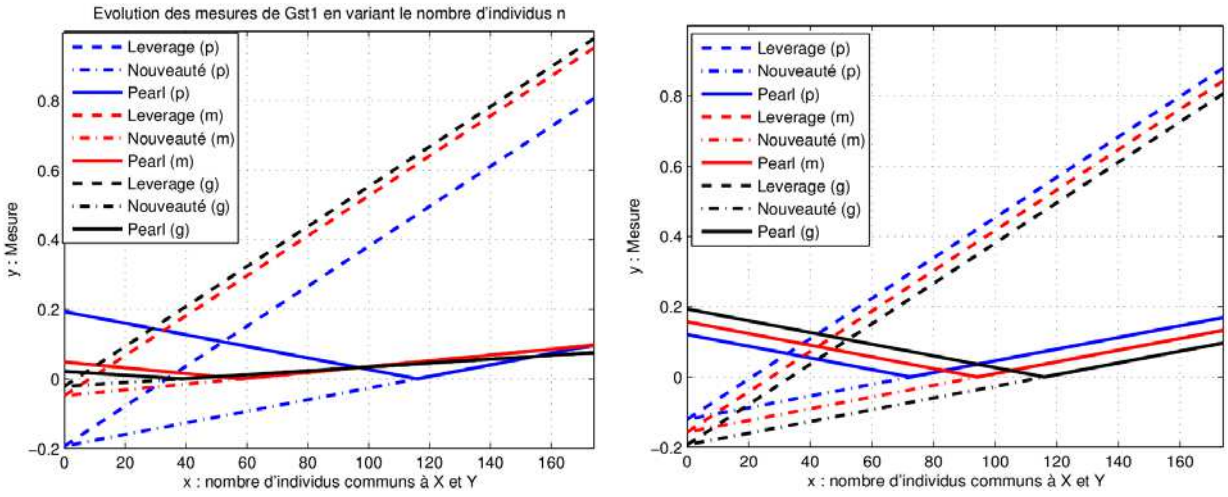

FIGURE *5.6: Évolution des mesures de G_{St1} en fonction du nombre d'exemples en variant le nombre total d'individus n (gauche) et la taille du conséquent (droite).*

pour vérifier les deux autres mesures de ce groupe.

– *Étape c* : Pour expliquer la présence des mesures *Leverage* et *Pearl* avec le couple {*Nouveauté, Piatetsky-shapiro*} et vérifier si elles ordonnent les règles de la même façon, nous traçons dans la *figure 5.6* les courbes d'évolution des mesures {*Leverage, Pearl et Nouveauté*} en fonction du nombre d'exemples. À l'issue de cette *figure 5.6* (*gauche*), nous illustrons la première *table 5.8* qui décrit les valeurs des trois mesures pour les deux cas d'exemple suivants : $n_{X_1Y_1} = 60$ et $n_{X_2Y_2} = 120$, et c'est pour chaque type de base (*p : petite, m : moyenne et g : grande*) tout en variant le nombre d'individus total (*figure 5.4*). À partir de cette table, nous remarquons que les mesures *Leverage* et *Nouveauté* ordonnent les règles de la même manière quelque soit le type de la base, et ceci en considérant la relation d'ordre requise entre ces mesures :

  Étant donnés $n_{X_1Y_1} = 60$ et $n_{X_2Y_2} = 120$, nous avons $Leverage(X_1 \rightarrow Y_1) = 0.18 \leq Leverage(X_2 \rightarrow Y_2) = 0.45 \quad et \quad Nouveauté(X_1 \rightarrow Y_1) = -0.5 \leq Nouveauté(X_2 \rightarrow Y_2) = 0$

  La mesure *Pearl* vérifie cette relation pour chaque type de base de données, à l'exception des bases de petite taille, i.e., les bases denses. Nous trouvons que $Pearl(X_1 \rightarrow Y_1) = 0.1 \geq Pearl(X_2 \rightarrow Y_2) = 0$.

  La *figure 5.6* (*droite*), présente les courbes d'évolution des trois mesures de ce groupe en fonction du nombre d'exemples et ceci en variant cette fois l'ensemble T_Y (*figure 5.5*). Cette figure vient confirmer le regroupement des 3 mesures (*3 courbes linéaires parallèles*) et montre que ces dernières ordonnent les règles de manière semblable.

Densité-Base	n_{XY}	**Mesure**		
–	–	Leverage	Nouveauté	Pearl
petite	60	0.18	-0.5	**0.1**
	120	0.45	0	**0**
moyenne	60	0.3	-0.9	-0.9
	120	0.62	0.03	0.03
grande	60	0.33	0.01	0.01
	120	0.63	0.027	0.027

TABLE *5.8: Valeurs que prennent les mesures de G_{St1} en variant le nombre d'exemples n_{XY} et le nombre d'individus total.*

Groupe G_{St2} :

- *Étape a* : Ce groupe comprend 7 mesures d'intérêt (*Sebag*, *Taux d'exemples*, *Laplace*, *Ganascia*, *Risque relatif*, M_{GK} et *Facteur de certitude*). Certaines de ces mesures sont liées par une transformation monotone croissante de la *Confiance* :
  $Sebag = \frac{1}{\frac{1}{Confiance}-1}$, $Taux\ d'exemples = 2 - \frac{1}{Confiance} = 1 - \frac{1}{Sebag}$.
  D'autres sont des variantes de la *Confiance* :
  $Ganascia = 2 \times Confiance - 1$, $Risque\ relatif = \frac{Confiance}{P(Y/\overline{X})}$ et enfin $Laplace = \frac{Confiance \times (n \times p(XY)+1)}{n \times p(XY)+2Confiance}$ qui prend en compte le nombre de transactions T.
  Nous retrouvons aussi des mesures ayant des relations de la forme :
  $M_{GK} = \frac{p(\overline{Y})}{p(Y)} \times Facteurdecertitude$ et $Risque\ Relatif = \frac{P(\overline{Y})Facteurdecertitude+P(Y)}{P(Y/\overline{X})}$.
  Les liens mathématiques ainsi découverts (*voir table 2.5, page 73*) entre les mesures sont peu informatives sur la ressemblance du comportement de ces mesures. Il est alors indispensable de passer à l'*étape b*.
- *Étape b* : Les deux classifications que nous avons menées dans les deux *chapitres 3* et 4 confirment la catégorisation des mesures *Sebag*, *Taux d'exemples*, *Ganascia* et *Laplace*. Nous retrouvons aussi les travaux de [VLL04] et [HGB+07] qui ont pu vérifié la similarité des trois mesures *Sebag*, *Taux d'exemples* et *Laplace*. Afin de vérifier le comportement de ces mesures avec les mesures restantes *Risque relatif*, *Facteur de certitude* et M_{GK}, nous passons à l'étape prochaine.
- *Étape c* : Selon [LMVL08], les mesures qui sont liées par une transformation monotone croissante sont équivalentes dans la mesure où elles classent les règles de la même façon. C'est le cas des mesures *Sebag* et *Taux d'exemples* décrites dans l'*étape a*. Nous cherchons dans cette étape à vérifier si toutes les mesures de ce groupe ordonnent les règles de la même façon. Pour ce faire, nous étudions seulement les mesures qui n'ont pas été vérifiées dans l'étape précédente {*Risque relatif, Facteur de certitude,* M_{GK}} et les comparer avec l'une des mesures du sous-groupe validé (*e.g., Sebag*). Les deux

Densité-Base	n_{XY}	**Mesure**			
–	–	Risque relatif	Facteur de certitude	M_{GK}	Sebag
petite	60	0.4	-1.0	-0.5	0.5
	120	1.1	0.1	0.1	2.25
moyenne	60	1.05	0.0	0.0	0.5
	120	2.5	0.5	0.5	2.25
grande	60	1.6	0.2	0.2	0.5
	120	4.0	0.6	0.6	2.25

TABLE *5.9: Valeurs que prennent les mesures de G_{St2} en variant le nombre d'exemples n_{XY} et le nombre total d'individus.*

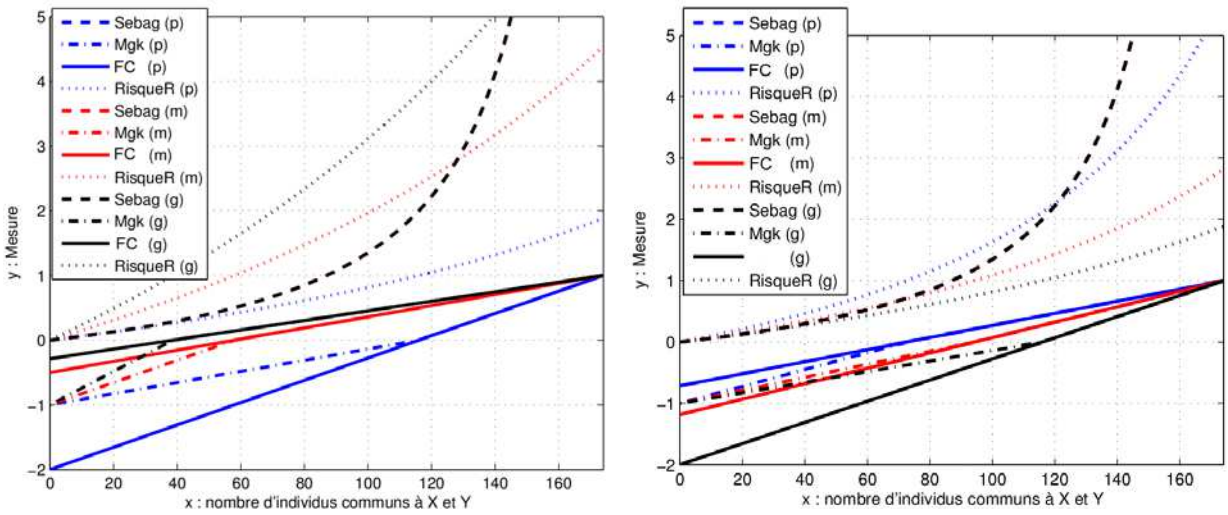

FIGURE *5.7: Évolution des mesures de G_{St2} en fonction du nombre d'exemples en variant le nombre total d'individus n (gauche) et la taille du conséquent (droite).*

graphes de la *figure 5.7* décrivent les courbes d'évolution de ces mesures en fonction du nombre d'exemples en variant respectivement le nombre total d'individus (*gauche*) et la taille du conséquent (*droite*). À partir de ces deux graphes, nous constatons que les mesures M_{GK} et *Facteur de certitude* sont équivalentes au niveau de la zone d'attraction, c'est-à-dire lorsque $P(Y/X) \geq P(Y)$, et nous confirmons le comportement invariable de la mesure *Sebag* en fonction de n et n_Y. La *table 5.9* illustre les différentes valeurs de ces 4 mesures pour les deux cas d'exemple suivants : $n_{X_1Y_1} = 60$ et $n_{X_2Y_2} = 120$; et c'est pour chaque type de base (*p : petite, m :moyenne et g :grande*) en variant le nombre d'individus total (*figure 5.4*). Ces valeurs montrent que toutes les mesures vérifient la relation d'ordre et classent les règles de la même manière quelque soit le type de la base de données, ce qui est aussi confirmé par la *figure 5.7* (*droite*) en variant le nombre du conséquent.

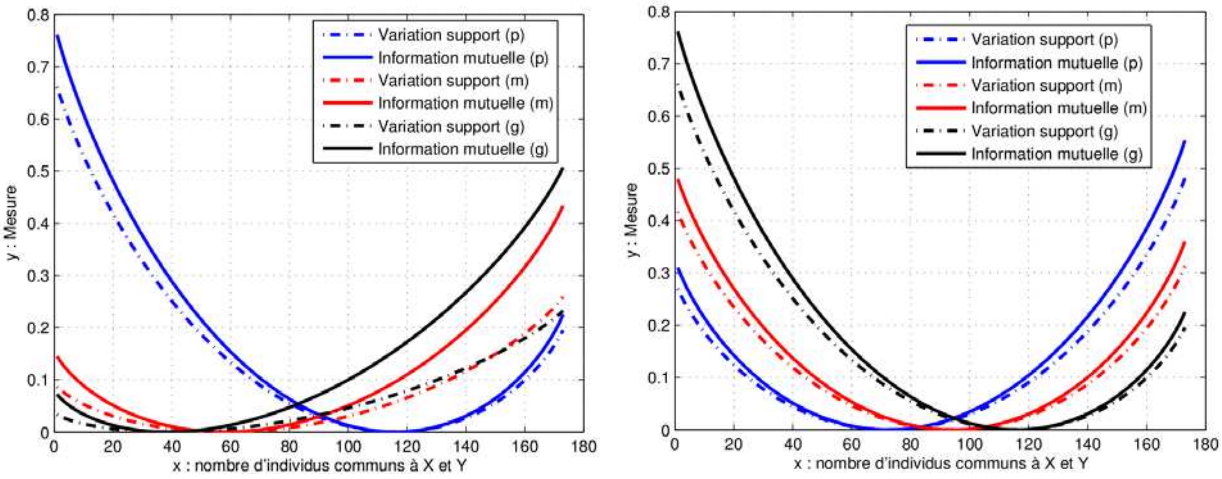

FIGURE *5.8: Évolution des mesures de G_{St3} en fonction du nombre d'exemples (gauche) et de la taille du conséquent (droite).*

Groupe G_{St3} :

- *Étape a* : Dans la *table 2.5, page 73*, nous découvrons des relations mathématiques de la forme $Information\ Mutuelle = \frac{Variation\ Support}{-p(X)log_2p(X)-p(\overline{X})log_2p(\overline{X})}$ entre les deux mesures de ce groupe. Certes, une telle liaison ne peut valider le comportement similaire de ces mesures.
- *Étape b* : Selon l'étude que nous avons menée dans le *chapitre 3*, nous découvrons que ce couple de mesures {*Information Mutuelle, Variation Support*} est vérifié par une seule méthode de classification, qui est la classification non hiérarchique des *k-moyennes* avec 17 propriétés communes (*cf. page 86*). L'étude menée dans le *chapitre 4 (voir page 138*) par la méthode avec recouvrement AFB confirme aussi ce regroupement. Toutefois, nous passons à l'étape suivante pour vérification.
- *Étape c* : Les deux graphes de la *figure 5.8* décrivent l'allure convexe des deux mesures *Information Mutuelle* et *Variation Support* en fonction du nombre d'exemples et montrent qu'elles ordonnent les règles de manière similaire en variant respectivement le nombre total d'individus et le nombre du conséquent.

Groupe G_{St4} :

- *Étape a* : Une relation logarithmique existe entre les deux mesures de ce groupe de la forme : $Gain\ Informationnel = log_2(Intérêt)$. Une telle relation confirme leur ressemblance.
- *Étape b* : Différents travaux de recherche vérifient ce regroupement de mesures, nous citons ceux de [LVML07] également, les catégories de mesures que nous avons identi-

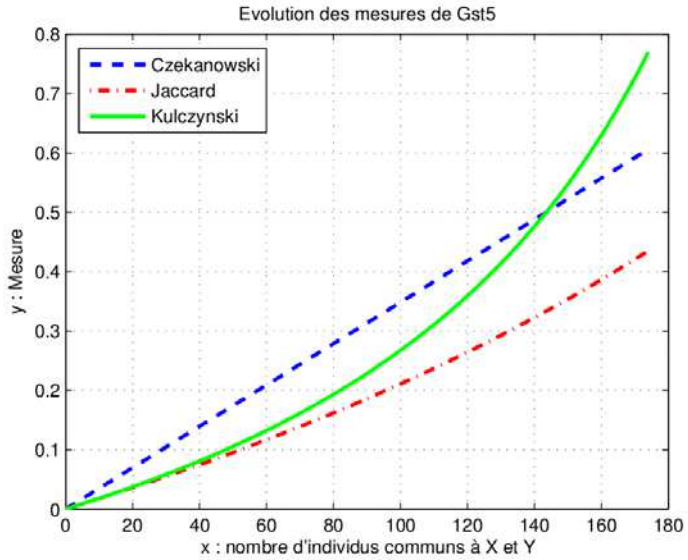

FIGURE *5.9: Évolution des mesures de G_{St5} en fonction du nombre d'exemples.*

fiées dans le *chapitre 3*, et finalement le travail de [Bra11] qui montre que ce couple de mesures évaluent les 6 propriétés qu'il a étudié de la même manière.

– *Étape c* : cette étape n'est plus indispensable puisque nous avons pu vérifier ce groupe stable à partir de l'*étape a* et confirmer dans l'*étape b*.

Groupe G_{St5} :

– *Étape a* : Des relations mathématiques de la forme : $Jaccard = \frac{Czekanowski-Dice}{2-Czekanowski-Dice}$ ou $Kulczynski = \frac{Jaccard}{1-Jaccard}$ existent entre les 3 mesures du groupe G_{St5} selon la *table 2.5, page 73*.

– *Étape b* : Ce groupe de mesures a été validé par l'étude formelle que nous avons réalisée dans le *chapitre 3* et par celle de Le Bras [Bra11]. Ces 3 mesures {*Jaccard, Czekanowski-Dice, Kulczynski*} ont montré à partir de ces deux travaux, qu'elles possèdent un comportement similaire en évaluant toutes les propriétés formelles étudiées de la même manière (*voir figure 5.9 qui illustre l'évolution de ces mesures en fonction du nombre d'exemples*). En outre, selon les travaux de *Lesot* et *Rifqi* [LR10] qui n'ont étudié que les deux mesures *Czekanowski-Dice* et *Jaccard*, les auteurs indiquent que celles-ci sont équivalentes.

– *Étape c* : Cette étape n'est pas nécessaire puisque nous avons pu confirmer ce groupe stable à partir des deux étapes précédentes.

Groupe G_{St6} :

– *Étape a* : Trois mesures d'intérêt {*Fiabilité négative, Confiance causale et Confiance-confirmée causale*} appartiennent à ce groupe. Aucune relation intéressante entre elles

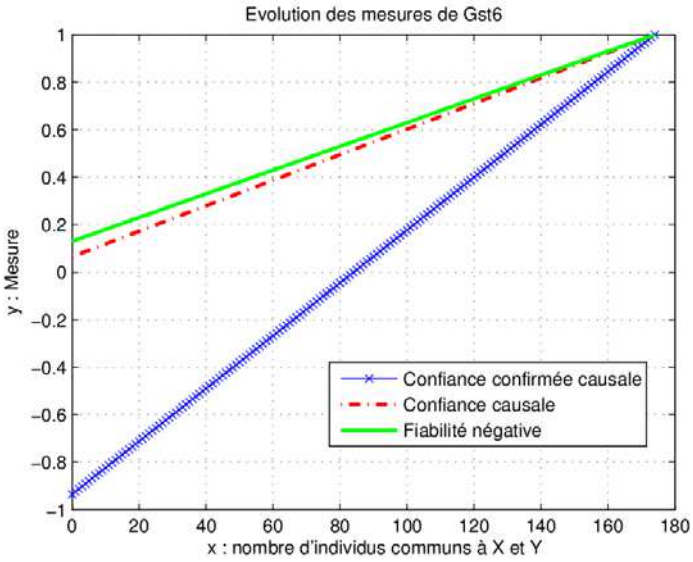

FIGURE *5.10: Évolution des mesures de G_{St6} en fonction du nombre d'exemples.*

n'est révélée. Nous passons ainsi à l'étape suivante.

- *Étape b* : Selon l'étude formelle que nous avons réalisée dans le *chapitre 3* (*cf. section 3.3.1, page 83*), nous expliquons que le groupe G_{St6} est aussi un groupe stable puisque ses trois mesures évaluent pareillement toutes les propriétés étudiées. La *figure 5.10* montre que ces trois mesures se rencontrent en un même point au niveau de l'implication logique. En outre, l'étude réalisée par *Huynh et al.* [HGB05a] sur la classification des mesures confirme le regroupement des mesures *Confiance causale* et *Confiance-confirmée causale*.
- *Étape c* : Ce groupe a été vérifié dans l'étape précédente (*étape b*). Par conséquent, nous n'étudions pas la relation d'ordre entre les mesures.

Groupe G_{St7} :

- *Étape a* : Ce groupe comprend les mesures statistiques qui sont basées sur l'entropie {*IPEE, IP3E, IIER, IIE*}, accompagnées des mesures *Support* et *Rappel*. Ces dernières sont liées par la formule : $Rappel = \frac{Support}{p(Y)}$.
- *Étape b* : L'étude formelle que nous avons réalisée dans le *chapitre 3* vérifie séparément les mesures (*Support, Rappel*) et les mesures (*IPEE, IP3E, IIE*). La mesure *IIER* rejoint les autres mesures statistiques de ce groupe selon la méthode ascendante hiérarchique et la méthode avec recouvrement du *chapitre 4*. Afin de s'assurer de ce regroupement de mesures, nous passons à l'étape suivante.
- *Étape c* : Pour étudier l'ordonnancement des règles par ces mesures, nous gardons une des mesures statistiques de ce groupe (*IIE*) plus les mesures *Support* et *Rappel*, et nous

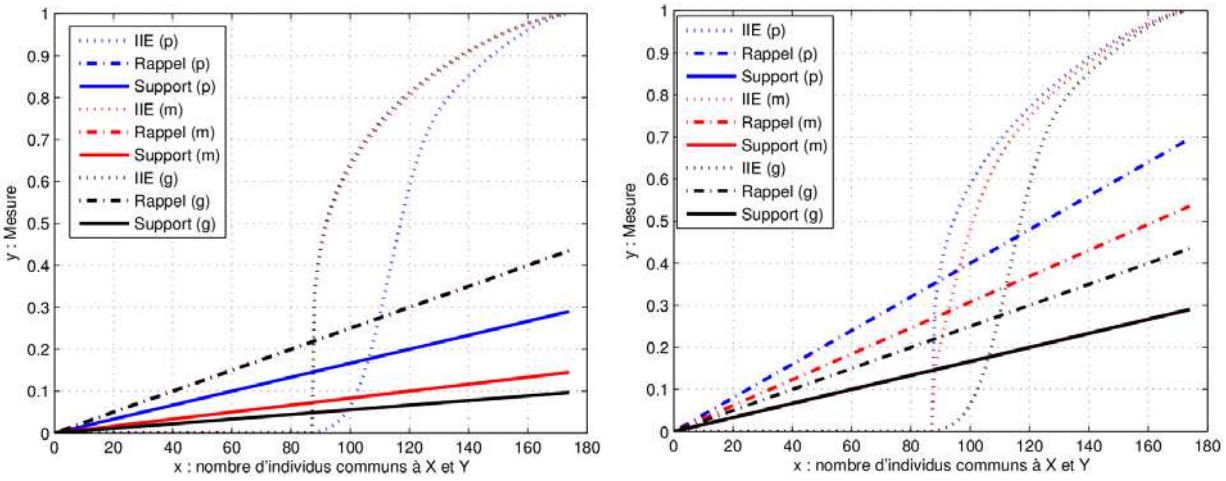

FIGURE *5.11: Évolution des mesures de G_{St7} en fonction du nombre d'exemples en variant le nombre d'individus total n (gauche) et le nombre du conséquent (droite).*

traçons leur courbes d'évolution en fonction du nombre d'exemples en variant tout d'abord le nombre total d'individus (*figure 5.11 (gauche)*) par la suite le nombre du conséquent (*figure 5.11 (droite)*). Selon le premier graphe de gauche de la *figure 5.11*, nous remarquons que la mesure *Rappel* est invariable selon le nombre d'individus total et que la relation d'ordre est vérifiée par ces trois mesures. Nous confirmons ceci à partir de la *table 5.10*, qui décrit les valeurs des 3 mesures pour ces deux cas de n_{XY} : $n_{X_1Y_1} = 60$ *et* $n_{X_2Y_2} = 120$ et ceci pour chaque type de base de données. Nous avons par exemple les relations suivantes :

Pour le cas des bases denses, $IIE(X_1 \rightarrow Y_1) = 0.00 \leq IIE(X_2 \rightarrow Y_2) = 0.60 \quad et \quad Support(X_1 \rightarrow Y_1) = 0.10 \leq Support(X_2 \rightarrow Y_2) = 0.2$.

Pour le cas des bases moyennes, nous trouvons $IIE(X_1 \rightarrow Y_1) = 0.00 \leq IIE(X_2 \rightarrow Y_2) = 0.81 \quad et \quad Support(X_1 \rightarrow Y_1) = 0.05 \leq Support(X_2 \rightarrow Y_2) = 0.10$.

Et finalement les bases éparses, $IIE(X_1 \rightarrow Y_1) = 0.00 \leq IIE(X_2 \rightarrow Y_2) = 0.82 \quad et \quad Support(X_1 \rightarrow Y_1) = 0.03 \leq Support(X_2 \rightarrow Y_2) = 0.06$.

Le deuxième graphe de droite de la *figure 5.11* vient ainsi confirmer les résultats révélés par la variation du nombre total d'individus et montre le comportement identique de ces mesures qui ordonnent de la même manière les règles quelque soit le type de la base.

Groupe G_{St8} :

– *Étape a* : Les deux mesures *Facteur bayésien* et *Conviction* de ce dernier groupe stable sont liées par la formule suivante : $Facteur\ Bayésien = Conviction \times Intérêt$. Nous retrouvons que ce couple de mesures est en corrélation avec la mesure *Intérêt*. Mais

Densité-Base	n_{XY}	Mesure		
–	–	IIE	Rappel	Support
petite	60	0.00	0.15	0.10
	120	0.60	0.30	0.20
moyenne	60	0.00	0.15	0.05
	120	0.81	0.30	0.10
grande	60	0.00	0.15	0.03
	120	0.82	0.30	0.06

TABLE *5.10: Valeurs que prennent les mesures de G_{St7} en variant le nombre d'exemples n_{XY} et le nombre d'individus total.*

l'étude empirique ne valide pas cette liaison. Ainsi, nous allons chercher à partir de l'étape suivante à expliquer le regroupement de ces mesures.

– *Étape b* : Plusieurs chercheurs ont étudié les deux mesures de G_{St8} et ont pu les catégoriser ensemble. Nous trouvons ainsi les travaux de [LVML07] et [Bra11]. Ce dernier a étudié les deux mesures selon 6 propriétés et a montré qu'elles les évaluent toutes pareillement. Les deux classifications que nous avons menées dans les *chapitres 3* et *4*, *pages 89* et *138*, vérifient aussi le regroupement de ces mesures.
  Selon la *figure* 5.12, nous remarquons que l'allure de la courbe d'évolution de la mesure *Intérêt* (*linéaire*) en fonction du nombre d'individus communs à X et Y est différente de celles des mesures *Facteur bayésien* et *Conviction*. Ces deux dernières sont convexes privilégiant les contre-exemples.
– *Étape c* : Ce groupe stable a été vérifié dans l'étape précédente. Cette étape n'est donc pas nécessaire.

À l'issue de cette dernière étape de notre méthodologie expérimentale, nous avons pu identifier et interpréter les 8 groupes de mesures stables. Dans ce qui suit, nous cherchons à confronter les résultats obtenus empiriquement avec ceux de l'étude formelle réalisée dans le *chapitre 3*, *page 89*, dans le but de valider et de mettre en valeur notre précédente approche.

5.6 Confrontation avec les 7 classes de mesures

Nous cherchons dans cette section à vérifier les 7 classes de mesures obtenues par l'étude formelle menée dans le *chapitre 3, page 89*, et à confronter ces classes avec les résultats de l'étude expérimentale décrite dans la section précédente.

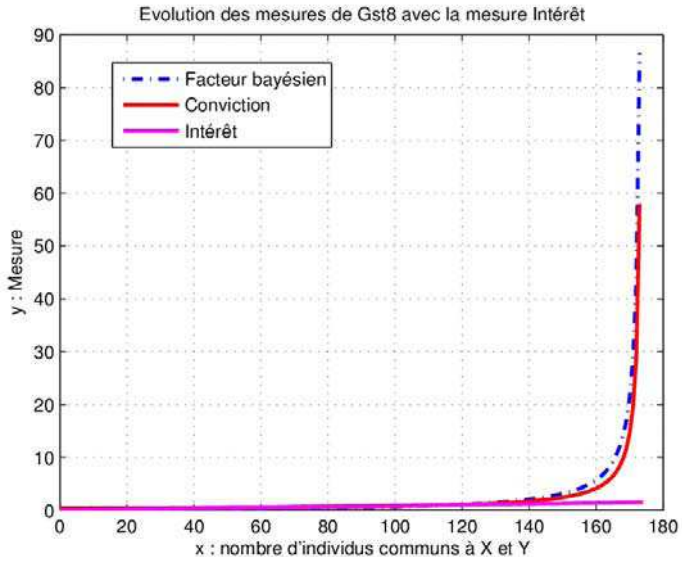

FIGURE *5.12: Évolution des mesures de G_{St8} en fonction du nombre d'exemples.*

5.6.1 Classe 1

La classe C_1 comprend le couple de mesures (*Indice de vraisemblance du lien (IVL), Intensité d'implication (II)*[6]). La matrice de similarité résultant de l'étude expérimentale, illustrée dans la *table 5.17*, indique des valeurs nulles (*colorées en bleu*) pour le degré de similarité I_S ainsi que l'écart-type σ pour ce couple de mesures (*II et IVL ne possèdent aucune règle commune*). Ainsi, la confrontation des deux résultats (*formels et empiriques*) révèle un désaccord qui ne nous permet pas de valider la classe C_1 empiriquement. Cependant, en regardant la matrice d'évaluation des mesures (*cf. chapitre 2, pages 70 et 71*), nous remarquons que ce couple de mesures évalue 19 propriétés identiquement, à l'exception des propriétés P_3 (*Mesure non symétrique*) et P_{15} (*Invariance en cas de dilatation de certains effectifs*). La mesure *IVL* est une mesure non symétrique, invariante en cas de dilatation de certains effectifs, contrairement à *II* qui est symétrique (P_3) et qui varie en cas de dilatation de certains effectifs (P_{15}), ce qui peut expliquer leur discordance.

Néanmoins, l'*indice de vraisemblance du lien* montre une ressemblance aux mesures statistiques appartenant à la classe C_2 pour le cas des bases "*CHESS*" et "*CONNECT*". Toutes ces mesures (*statistiques et IVL*), y compris l'*intensité d'implication* représentent les indices de la famille de l'indice de vraisemblance du lien [Ler70]. Certes, nous signalons l'absence de l'*intensité d'implication* dans ce groupe de mesures selon l'étude empirique qui peut-être due à sa variance dans le cas de dilatation de certains effectifs. Cette propriété (P_{15}) qui est non partagée ni avec *IVL* ni avec les mesures appartenant à la classe C_2, et qui semble être empi-

6. Nous considérons la loi binomiale pour cette mesure

Mesure	IPEE	IP3E	IIE	IIER
IPEE	1,00			
IP3E	1,00	1,00		
IIE	1,00	1,00	1,00	
IIER	1,00	1,00	1,00	1,00
I_S	1,00	1,00	1,00	1,00
σ	0,00	0,00	0,00	0,00

TABLE *5.11: Matrice de similarité de la classe* C_2.

riquement déterministe quand à la nature des règles extraites.

5.6.2 Classe 2

Toutes les mesures appartenant à la classe C_2 sont incluses dans le groupe stable G_{St7}, à l'exception de la mesure *Indice probabiliste discriminant* (*IPD*) que nous n'avons pas étudié empiriquement suite à la complexité de ce dernier. Ainsi, les mesures *IP3E, IPEE, IIE*, y compris la mesure *IIER* (*IIER appartient à* C_2 *selon la méthode hiérarchique*), possèdent un comportement identique puisqu'elles forment un groupe de mesures stables. La *table 5.11* est un extrait de la matrice de similarité, illustrée dans la *table 5.17*, mais représente essentiellement la classe C_2. Selon cette table, nous n'avons que des valeurs maximales de $I_S = 1$ et des valeurs nulles de l'écart-type, et ce pour tous les couples de mesures de C_2. Par conséquent, la classe C_2 est vérifiée empiriquement.

5.6.3 Classe 3

La classe C_3 comprend 7 mesures d'intérêt (*Indice d'implication, Gini, Dépendence, J-mesure, Information mutuelle, Prévalence* et *Couverture*) possédant empiriquement des comportements différents. Nous avons 4 mesures (*Indice d'implication, Gini, Prévalence* et *Couverture*) qui appartiennent à la *catégorie 2* (*étape 4 de la section 5.2, page 148*) pour leur comportement dissemblable et 2 mesures (*Dépendence, J-mesure*) qui appartiennent à la *catégorie 3* pour leur comportement indéterminé. Nous catégorisons ces mesures selon leurs valeurs de I_S et σ, décrites dans les *tables 5.16* et *5.17*.

La différence des résultats empiriques révélés par les mesures de la classe C_3 nous incite à revoir leur matrice caractéristique correspondante selon les propriétés (*chapitre 2, pages 70 et 71*). Cette dernière montre que toutes les mesures appartenant à C_3 sont non symétriques et ne vérifient pas la majorité des propriétés (P_6, P_7, P_8, P_{10}, P_{11}, P_{13}, P_{15}, P_{16}, P_{17}, P_{19}) étudiées.

Cependant, nous remarquons la présence de mesures additionnelles (*Variation support,*

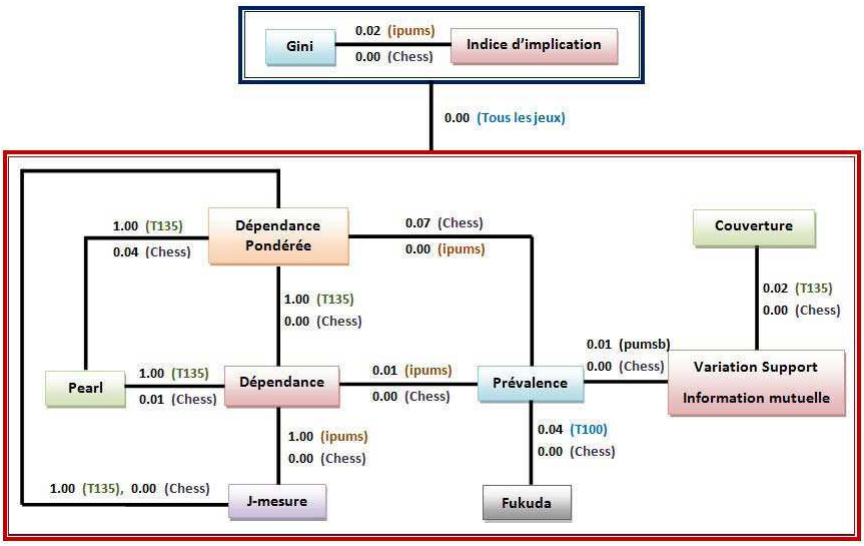

FIGURE *5.13: Pourcentage de règles communes entre les mesures de C_3, majoritairement faible, selon la base de données sélectionnée.*

Pearl, Dépendance pondérée, Fukuda) qui appartiennent à la classe C_3 selon l'une des méthodes de classification sans recouvrement appliquées dans le *chapitre 3* (*voir page 89*). En considérant ces mesures, nous retrouvons le couple (*Variation support, Information mutuelle*) qui forme le groupe stable G_{St3} ainsi que nous remarquons une similarité entre les mesures (*Dépendence, J-mesure*) de la classe C_3 et les mesures additionnelles (*Pearl, Dépendance pondérée, Fukuda*), que nous décrivons dans la *figure 5.13*. Par exemple, nous avons La mesure *Dépendance* qui est proche de la mesure *Pearl* avec un taux de ressemblance $\tau_{B_6} = 1,00$ au cas où la base sélectionnée est "*T135L23N60*" et de $\tau_{B_1} = 0,01$ pour la base "CHESS".

La mesure *Dépendance* est aussi proche de plusieurs autres mesures (*Leverage*, *Spécificité*, *Confiance causale*, *Fiabilité négative* et *Confiance confirmée causale*), notamment de la classe C_4 pour un taux de ressemblance τ compris entre $0,14$ (*Dépendance-Fiabilité négative pour la base "IPUMS"*) et $1,00$ (*principalement résultant de la base "T135L23N60"*). Ce rapprochement peut être vérifié par le dendogramme décrit dans le *chapitre 3, page 85*, qui nous présente le regroupement des mesures déjà citées lorsque nous le coupons à un niveau de hiérarchie plus élevé.

5.6.4 Classe 4

La classe C_4 comprend les 14 mesures d'intérêt suivantes : *Confiance causale*, *Fiabilité négative*, *Confiance confirmée causale*, *Czekanowski-Dice*, *Kulczynski*, *Jaccard*, *Leverage*, *Spécificité*, *Dépendance causale*, *Précision*, *Cosinus*, *Confirmation causale*, *Rappel*, *Support*. Selon

l'étude empirique, cette classe contient trois groupes stables : G_{St6} qui comprend les mesures (*Confiance causale*, *Fiabilité négative* et *Confiance confirmée causale*), G_{St5} qui contient les trois mesures (*Kulczynski, Jaccard* et *Czekanowski-Dice*) et deux mesures de G_{St7} qui sont (*Rappel et Support*). La *table 5.12* présente le taux de ressemblance entre les différents couples de mesures appartenant à la classe C_4, y compris la mesure *VT100* qui lui appartient selon la méthode hiérarchique. Il met aussi en exergue le comportement indéterminé (*en fonction des bases de données*) de certaines de ces mesures, que nous classons par conséquent dans la *Catégorie 3*. Par exemple, nous trouvons que le groupe G_{St6} est proche de *Leverage* avec un taux de ressemblance $\tau_{B_6} = 1,00$ pour le cas de la base "*T135L23N60*" et de $\tau_{B_3} = 0,18$ pour le cas de la base "IPUMS".

Concernant la mesure *Précision*, l'étude expérimentale montre qu'elle est assez distante des autres mesures de la même classe C_4. La matrice d'évaluation (*chapitre 2*, page 71) montre que cette dernière est la seule dans son groupe qui vérifie la propriété P_{18} (*Égalité entre les règles* $X \rightarrow Y$ *et* $\bar{X} \rightarrow \bar{Y}$), ce qui peut expliquer cette dissemblance.

À l'issue de cette étude, nous constatons que la classe C_4 est partiellement vérifiée puisqu'elle contient des mesures ayant un comportement identique et d'autres qui ont un comportement indéterminé, en fonction de la base de données.

5.6.5 Classe 5

La classe C_5 contient 7 mesures d'intérêt, dont 4 mesures (*Sebag, Ganascia, Taux d'exemples* et *Laplace*) forment le groupe stable G_{St2}. Ces mesures stables sont aussi proches des mesures *Confiance* et *Moindre contradiction* pour différentes valeurs du taux de ressemblance telles que décrites dans la *figure 5.14*. Seule la mesure *Confirmation Descriptive* possède un comportement différent puisque son degré de similarité I_S est toujours nul avec les autres mesures de la classe C_5 (*cette discordance n'est pas expliquée par la matrice d'évaluation du chapitre 2, page 70*).

Toutes ces mesures, à l'exception de la mesure *Confirmation Descriptive* partagent selon l'étude formelle (*chapitre 2, pages 70 et 71*) 14 propriétés communes ($P_3 = 1$, $P_4 = 1$, $P_6 = 1$, $P_7 = 0$, $P_8 = 0$, $P_9 = 0$, $P_{11} = 1$, $P_{12} = 0$, $P_{13} = 0$, $P_{15} = 0$, $P_{16} = 0$, $P_{18} = 0$, $P_{19} = 0$ et $P_{20} = 0$).

5.6.6 Classe 6

La classe C_6 contient les 5 mesures d'intérêt suivantes : *Zhang*, M_{GK}, *Q de Yule*, *Y de Yule* et *Goodman*. L'étude empirique montre que les 3 dernières mesures sont différentes des 2 autres mesures de la classe C_6 pour de très faibles valeurs de I_S et σ. Ces trois mesures

Mesures / Bases		B1	B2	B3	B4	B5	B6
G_{St6}	Dépendance pondérée	0,00	0,00	0,00	0,00	0,57	0,00
	Leverage	0,00	0,01	0,18	0,04	0,00	1,00
	Spécificité	0,00	0,00	0,14	0,10	0,00	1,00
	Cosinus	0,00	0,00	0,00	0,00	0,00	0,00
	Confirmation causale	0,00	0,00	0,00	0,04	0,00	0,00
	Rappel	0,00	0,00	0,00	0,00	0,00	0,00
	G_{St5}	0,00	0,00	0,00	0,00	0,00	0,00
	VT100	0,01	0,00	0,00	0,00	0,00	1,00
Dépendance pondérée	Leverage	0,00	0,00	0,00	0,00	0,00	0,00
	Spécificité	0,00	0,00	0,00	0,00	0,00	0,00
	Cosinus	0,00	0,00	0,02	0,02	0,00	0,02
	Confirmation causale	0,00	0,00	0,02	0,00	0,00	0,02
	Rappel	0,00	0,00	0,00	0,80	0,00	0,00
	G_{St5}	0,00	0,00	0,02	0,02	0,00	0,02
	VT100	0,00	0,01	0,02	0,02	0,00	0,00
Leverage	Spécificité	0,00	0,00	0,78	0,08	0,00	1,00
	Cosinus	0,00	0,00	0,00	0,00	0,00	0,00
	Confirmation causale	0,00	0,00	0,00	0,00	0,00	0,00
	Rappel	0,00	0,00	0,00	0,00	0,00	0,00
	G_{St5}	0,00	0,00	0,00	0,00	0,00	0,00
	VT100	0,00	0,00	0,00	0,00	0,00	1,00
Spécificité	Cosinus	0,00	0,00	0,00	0,01	0,00	0,00
	Confirmation causale	0,00	0,00	0,00	0,00	0,00	0,00
	Rappel	0,00	0,00	0,00	0,00	0,00	0,00
	G_{St5}	0,00	0,00	0,00	0,01	0,00	0,00
	VT100	0,00	0,00	0,00	0,00	0,02	1,00
Cosinus	Confirmation causale	0,02	0,01	1,00	0,14	0,02	1,00
	Rappel	0,35	0,22	0,00	0,00	0,34	0,00
	G_{St5}	0,38	0,49	1,00	0,85	0,39	1,00
	VT100	0,00	0,00	1,00	0,02	0,00	0,00
Confirmation causale	Rappel	0,02	0,01	0,00	0,02	0,02	0,00
	G_{St5}	0,01	0,01	1,00	0,14	0,02	1,00
	VT100	0,00	0,00	1,00	0,02	0,00	0,00
Rappel	G_{St5}	0,58	0,36	0,00	0,00	0,71	0,00
	VT100	0,00	0,00	0,00	0,00	0,00	0,00
G_{St5}	VT100	0,00	0,00	1,00	0,02	0,00	0,00

TABLE *5.12: Pourcentage de règles communes entre les mesures de C_4, qui est parfois faible et parfois élevé, selon le jeu de données sélectionné.*

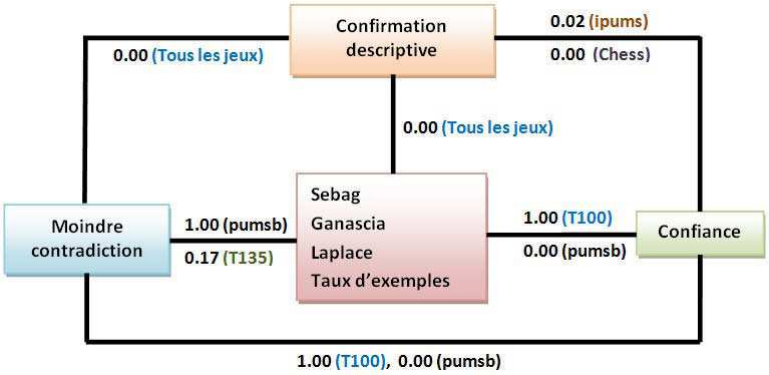

FIGURE *5.14: Pourcentage de règles communes entre les mesures de C_5 selon le jeu de données sélectionné.*

possèdent selon la matrice d'évaluation (*chapitre 2, pages 70 et 71*) un caractère symétrique, ce qui n'est pas le cas des mesures M_{GK} et *Zhang*. Cette différence pourrait être probablement expliquée par la dissemblance dans la nature des règles extraites entre ces mesures. De plus, nous remarquons que les mesures M_{GK} et *Zhang* sont d'une part très proches selon les bases de données (*CONNECT, IPUMS, PUMSB, T100L10N40, T135L23N60*) pour une valeur élevée du taux de ressemblance $\tau_{B_{2,3,4,5,6}} = 1,00$ et d'autre part, elles sont complètement différentes selon la base de données "*CHESS*" uniquement, pour $\tau_{B_1} = 0,00$.

En parcourant la matrice d'évaluation des mesures (*chapitre 2*), nous nous apercevons que *Y de Yule* est la seule des mesures de C_6 qui possède une allure *convexe* et qui ne tolère pas les premiers contre-exemples. Néanmoins, cette dernière est assez proche des mesures (*Information mutuelle et Variation support*) du groupe stable G_{St3} avec un taux de ressemblance $\tau_{B_1} = 0,64$ (*CHESS*), $\tau_{B_{2,3,6}} = 1,00$, (*CONNECT, IPUMS, T135L23N60*), $\tau_{B_4} = 0,20$ (*PUMSB*) et $\tau_{B_5} = 0,16$ (*T100L10N40*). Nous pouvons expliquer ces valeurs par le comportement similaire des mesures quant à l'évaluation des 9 propriétés suivantes ($P_8 = 0$, $P_{11} = 0$, $P_{18} = 1$, $P_7 = 0$, $P_{19} = 0$, $P_{20} = 0$, $P_{21} = 1$ et $P_{14.3} = 1$).

5.6.7 Classe 7

La classe C_7 contient 10 mesures d'intérêt, à partir desquelles trois groupes stables sont identifiés empiriquement : (i) G_{St4} qui comprend les mesures *Intérêt* et *Gain informationnel*, (ii) G_{St8} qui comprend le couple de mesures *Conviction* et *Facteur bayésien* et (iii) G_{St2} qui contient les mesures *Risque relatif* et *Facteur de certitude* (*ce groupe contient d'autres mesures de C_5 et C_6*). Vu ces trois groupes de mesures stables, nous allons chercher s'il existe une liaison entre eux et entre les mesures restantes de cette même classe C_7. Pour ce faire, nous visualisons

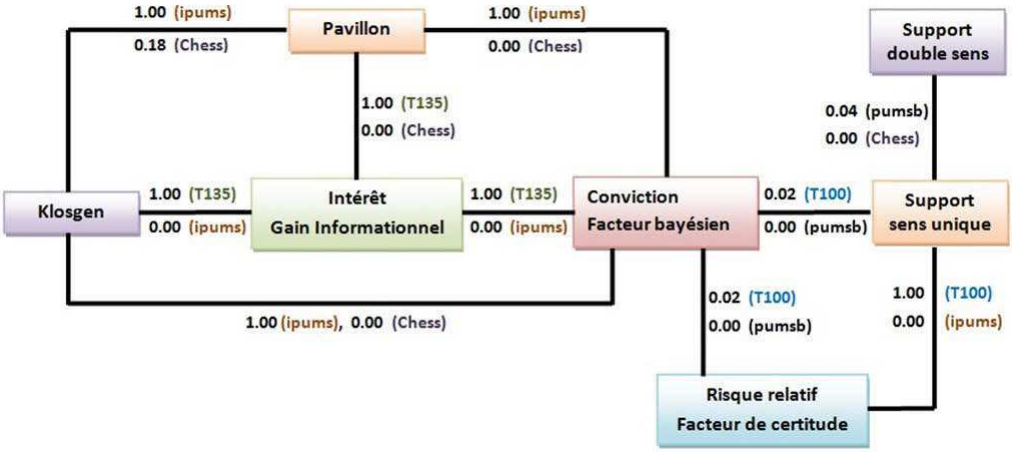

FIGURE *5.15: Pourcentage de règles communes entre les mesures de C_7, qui peut être à la fois faible et élevé, selon le jeu de données sélectionné.*

les p matrices de similarité révélées par l'étude empirique, à partir desquelles nous découvrons que le taux de similarité entre les différentes mesures de ce groupe varie entre $0,00$ et $1,00$. Autrement dit, certaines mesures de C_7 ne possèdent aucune règle commune quelque soit la base de données, telles que {*Risque relatif, Facteur de certitude, Klosgen, Support double sens*}, {*Pavillon, Support sens unique, Support double sens*}, tandis que d'autres se ressemblent avec des taux variables, en fonction de la base de données sélectionnée. La *figure 5.15* décrit les relations existantes entre les différentes mesures de la classe C_7. Nous mentionnons que les mesures qui ne se ressemblent pas empiriquement pour toutes les bases de données (*i.e., avec* $\tau = 0.00$) ne sont pas liées par un trait dans la *figure 5.15*.

5.6.8 Les mesures restantes

Dans le *chapitre 3*, nous avons indiqué qu'aucun consensus n'a été révélé pour ces mesures (*Piatetsky-shapiro, Nouveauté, Corrélation , Cohen, Force collective, VT100, Ratio des chances, IIER, Variation support, Pearl, Fukuda et dépendance pondérée*). De ce fait, nous allons chercher si nous pouvons les classer empiriquement. Tout d'abord, nous retrouvons que les mesures *Piatetsky-shapiro*, *Corrélation* et *Nouveauté* appartiennent à la classe C_6 selon la méthode de classification hiérarchique uniquement, voir *pages 85* et *89* du *chapitre 3*. Deux parmi ces mesures (*Piatetsky-shapiro, Nouveauté*) possèdent un comportement identique puisqu'elles appartiennent au groupe stable G_{St1}, mais empiriquement ne révèlent aucune liaison avec les mesures de la classe C_6.

Quand à la mesure *Corrélation*, elle semble être plus proche des mesures appartenant à la classe C_4 selon les matrices de similarité (*illustrées dans les tables 5.16 et 5.17*) qu'aux me-

Mesures / Bases		B1	B2	B3	B4	B5	B6
Piatetsky-shapiro	Nouveauté	$1,00$	$1,00$	$1,00$	$1,00$	$1,00$	$1,00$
	Force collective	$0,00$	$0,00$	$0,72$	$0,12$	$0,20$	$1,00$
	Ratio des chances	$0,00$	$0,00$	$0,72$	$0,80$	$0,00$	$1,00$
	Cohen	$0,00$	$0,00$	$0,00$	$0,80$	$0,00$	$0,40$
	VT100	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$1,00$
	Corrélation	$0,00$	$0,00$	$0,00$	$0,80$	$0,00$	$0,00$
Force collective	Ratio des chances	$0,00$	$0,54$	$1,00$	$0,66$	$0,00$	$1,00$
	Cohen	$0,00$	$1,00$	$0,00$	$0,56$	$0,16$	$0,00$
	VT100	$0,00$	$0,00$	$0,00$	$0,00$	$0,00$	$1,00$
	Corrélation	$0,00$	$0,54$	$0,00$	$0,62$	$0,00$	$0,00$
Ratio des chances	Cohen	$0,00$	$0,54$	$0,00$	$0,64$	$0,00$	$0,00$
	VT100	$0,38$	$0,00$	$0,00$	$0,00$	$0,80$	$1,00$
	Corrélation	$0,00$	$0,72$	$0,00$	$0,64$	$0,00$	$0,20$
Cohen	VT100	$0,00$	$0,00$	$1,00$	$0,20$	$0,00$	$0,00$
	Corrélation	$0,12$	$0,54$	$1,00$	$0,90$	$0,20$	$1,00$
VT100	Corrélation	$0,00$	$0,00$	$1,00$	$0,20$	$0,80$	$0,00$

TABLE *5.13: Pourcentage de règles communes entre les mesures restantes en fonction du jeu de données.*

sures de la classe C_6. La méthode de classification non hiérarchique des *k-moyennes* appliquée dans le *chapitre 3*, *page 86*, a montré que les trois mesures (*Piatetsky-shapiro*, *Corrélation* et *Nouveauté*) y compris *VT100*, *Force collective*, *Cohen* et *Ratio des chances* forment ensemble une nouvelle classe de mesures nommée "classe G_{p8}". L'étude expérimentale vient prouver ce regroupement des mesures pour un taux de ressemblance τ_{B_k} compris entre $0,00$ et $1,00$. La *table 5.13* expose en terme de valeurs le taux de ressemblance entre ces mesures pour chacun des jeux de données[7] étudiés et montre qu'il y a une liaison assez variable (*faible-moyenne-forte*) entre toutes ces mesures. Cette variance dépend en effet des mesures ainsi que des bases de données sélectionnées par l'utilisateur.

À l'issue de la confrontation des résultats de l'étude empirique avec les 7 classes de mesures obtenues par l'étude formelle menée dans le *chapitre 3, page 89*, nous présentons dans ce qui suit un tableau qui catégorise les mesures d'intérêt selon leur comportement empirique.

5.6.9 Catégorisation des mesures selon leur comportement empirique

Dans cette section, les *tables 5.14* et *5.15* décrivent le comportement des mesures étudiées selon l'étude empirique. Trois catégories de mesures sont identifiées dans l'*étape 4* de notre méthodologie expérimentale (*section 5.2*) : (**Cat1**) les mesures au comportement similaire,

7. *B1 = "CHESS", B2 = "CONNECT", B3 = "IPUMS", B4 = "PUMSB", B5 = "T100" et B6 = "T135"*

(**Cat2**) les mesures au comportement différent et (**Cat3**) les mesures au comportement indéterminé, en fonction des bases de données. Toutefois, au fur et à mesure que nous catégorisons les différentes mesures d'intérêt selon leur comportement empirique, nous vérifions aussi les 7 classes révélées par l'étude formelle, *page 89*.

Cette vérification se réfère alors aux notations suivantes : (a) "**V**" qui signifie "*Vérifiée*", c'est le cas où toutes les mesures de cette classe possèdent un comportement similaire, (b) "**VP**" signifie "*Vérifiée Partiellement*" i.e., la classe en question peut contenir ou bien à la fois des mesures qui appartiennent à la catégorie 1 (*Cat1*) et catégorie 3 (*Cat3*), ou bien contenir seulement des mesures de la catégorie 3. Enfin, (c) "**NV**" qui signifie "*Non Vérifiée*", c'est que les mesures de cette classe appartiennent à la catégorie 2, ce qui ne nous permet pas de la valider.

En visualisant les *tables 5.14* et *5.15*, nous nous apercevons qu'une classe est vérifiée (C_2), 2 classes ne sont pas vérifiées (C_1, C_3) et que 5 classes sont partiellement vérifiées (C_4, C_5, C_6, C_7 *et* G_{p_8}). La dernière classe G_{p_8} a été identifiée par la méthode des *k-moyennes* dans le *chapitre 3, page 86*. Ainsi, vu que la majorité des classes sont partiellement vérifiées, ceci montre que le comportement d'une mesure sur un jeu de données ne peut être généralisé à toutes les situations.

Nous notons que dans la *table 5.14*, certaines mesures seront suivies des notations (p) ou (c), qui désignent les méthodes de classification non supervisée appliquées, et signifient respectivement la méthode de classification des *k-moyennes* et la méthode de classification hiérarchique. Ces notations montrent la méthode de classification qui a regroupé la mesure en question avec les mesures d'une classe C_i (*i= 1..7*).

Dans ce qui suit, nous reprenons certains travaux [Vai06], [HGB+07], [Bra11] réalisés sur la classification des mesures et nous les comparons avec les résultats empiriques que nous avons obtenus.

5.7 Comparaison avec les autres travaux

Cette section s'intéresse à une étude comparative avec des travaux existants. Nous commençons par les travaux de Vaillant [Vai06].

5.7.1 Comparaison avec le travail de Vaillant

Une série d'expérimentations du comportement de 20 mesures d'intérêt sur 10 bases de données issues de l'UCI Repository [8] a été réalisée par [Vai06]. Cette étude a conduit à

8. http ://ftp.ics.uci.edu/pub/machine-learning-databases

l'identification de 4 classes principales de mesures d'intérêt :

- Cl_{v1} : {*Piatetsky, Cohen, Gain informationnel, Intérêt, Corrélation, Pavillon, Indice d'implication, IPD, intensité d'implication}* ;
- Cl_{v2} : {*IIE}* ;
- Cl_{v3} : {*Facteur de certitude, Zhang, Conviction, Confiance, Sebag, Laplace, Facteur bayésien, Taux d'exemples}* ;
- Cl_{v4} : {*Support, Moindre contradiction}*.

Ainsi, nous cherchons à travers nos différentes matrices de similarité (*illustrées dans les tables 5.16 et 5.17*) de mesures à vérifier ces 4 groupes de mesures par l'étude des valeurs des degrés de similarité. Nous remarquons que :

- La première classe Cl_{v1} présente une similarité entre ses mesures pour $I_S \in [0.00, 1.00]$. Nous remarquons la présence du groupe stable G_{St4} dans Cl_{v1} (*section* 5.6.8 présente en détail les relations de similarité existante entre les mesures *Piatetsky, Cohen* et *Corrélation*) ;
- la deuxième classe Cl_{v2} ne présente qu'une seule mesure ;
- la troisième classe Cl_{v3} est proche de notre groupe stable G_{St2} ainsi que de la classe C_5 puisqu'elles possèdent 4 mesures communes {*Facteur de certitude, Confiance, Sebag, Laplace, Taux d'exemples}* ;
- la dernière classe Cl_{v4} regroupe *2* mesures qui exposent une forte discordance selon les résultats empiriques puisqu'elles n'extraient aucune règle commune.

Du fait de ces observations, nous mentionnons que nous avons détecté à la fois des similarités et des dissimilarités au niveau des groupements de mesures réalisés par Vaillant. Nous expliquons cette différence par son choix de petits jeux de données provenant de l'UCI Repository, sur lesquels l'auteur a réalisé ses expérimentations. Mais face au problème de stockage de grandes quantités de données qui ne cesse d'augmenter jour après jour, les résultats obtenus par Vaillant sur la classification des mesures d'intérêt semblent être peu solides vue la taille des bases sur lesquelles il a travaillé.

5.7.2 Comparaison avec le travail de Hyunh et al.

Une autre classification a été effectuée par Huynh et al. [HGB+07]. Ces auteurs ont étudié 36 mesures d'intérêt, dont 32 mesures communes, sur 2 jeux de données de nature opposée :

l'un fortement corrélé (*mushroom*) et l'autre faiblement corrélé (*base synthétique T5.I2.D10K*). Les auteurs présentent dans un premier temps une taxonomie des mesures selon les 2 critères suivants :

1. *le sujet* : la déviation de l'indépendance ou de l'équilibre ;
2. *la nature* : descriptive ou statistique.

À partir de l'étude de ces deux paramètres particuliers sur les jeux de données, les 5 groupes de mesures suivants sont retenus. Nous mentionnons le taux de ressemblance pour chaque groupe identifié par [HGB+07] :

- Cl_{de} (*descriptive/déviation de l'équilibre*) : contient les mesures {*Confiance, Laplace, Sebag, Taux d'exemples, Confirmation descriptive, Confiance confirmée-descriptive, Moindre contradiction*} qui se ressemblent avec un taux compris entre $0,00$ et $1,00$. Certaines de ces mesures sont fortement indépendantes, contrairement à d'autres, e.g., nous retrouvons les mesures appartenant au groupe stable G_{St2} incluses dans Cl_{de} ;
- Cl_{di} (*descriptive/déviation de l'indépendance*) : ce groupe contient les mesures {*Corrélation, Intérêt, Facteur de certitude, Conviction, Dépendance, Pavillon, J-mesure, Gini, TIC, Force collective, Ratio des chances, Q de Yule, Y de Yule, Klosgen, Cohen*}. Les résultats de l'étude empirique vérifie ce regroupement de mesures (*à l'exception de le mesure TIC que nous n'avons pas étudié*) pour un taux de ressemblance τ_{ij} compris entre $0,00$ et $1,00$ selon les mesures et le jeu de données sélectionnés ;
- Cl_{se} (*statistique/déviation de l'équilibre*) : contient l'unique mesure {*IPEE}*. Cette mesure ressemble à la mesure *IIE* de la classe suivante Cl_{si} ;
- Cl_{si} (*statistique/déviation de l'indépendance*) : regroupe ces 5 mesures {*II, IIE, IIE2, Lerman, Règle d'intérêt*}. Seulement deux de ces mesures, *II* et *IIE* ont été étudiées empiriquement et nous avons trouvé qu'elles sont fortement indépendantes ;
- Cl_{o} (*autres*) : contient les mesures {*Support, Précision, Jaccard, Cosinus, Confiance causale, Confirmation causale, Confiance confirmée-causale, Dépendance causale*} qui appartiennent à la classe C_4, *page 89*. Comme le montre la *table 5.14*, ces mesures n'appartiennent pas toutes à une même catégorie, mais elles diffèrent l'une de l'autre.

Cette étude analytique des mesures proposée par [HGB+07] a mené aussi à la découverte de ces 5 groupes de mesures stables, où seules les mesures communes sont indiquées :

- Cl_{s1} : {*Confiance, Confiance causale, Confiance confirmée-causale, Ganascia, Laplace*} ;
- Cl_{s2} : {*Jaccard, Cosinus, Cohen, Corrélation*} ;
- Cl_{s3} : {*Confiance centrée, Dépendance, Dépendance causale*} ;

– Cl_{s4} : {*IIE*} ;
– Cl_{s5} : {*Q de Yule, Y de Yule*}.

La comparaison de nos résultats (*section 5.5*) avec ces 5 groupes de mesures montre une ressemblance dans nos travaux au niveau des 2 sous-groupes stables suivants : {*Confiance causale, Confiance-confirmée causale*} et {*Ganascia, Laplace*}. La différence existante avec les autres groupes stables peut-être expliquée par le nombre de jeux de données choisis (*2 jeux*) par les auteurs. En effet, ce petit nombre ne nous permet pas de savoir si les mesures sur lesquelles nous sommes en désaccord dépendent du jeu de données analysé ou pas.

Dans ce qui suit, nous confrontons nos résultats avec ceux de Le Bras [Bra11].

5.7.3 Comparaison avec le travail de Le Bras

Le Bras [Bra11] cherche dans ses travaux à trouver des caractéristiques communes aux mesures objectives. Pour ce faire, il a étudié 42 mesures d'intérêt selon 6 critères opérationnels qu'il a proposé. Ces critères concernent d'une part la possibilité de calculer la robustesse, et d'autre part d'utiliser des algorithmes efficaces. Les 6 critères sont : $Br_{1.1}$ (*Mesure plane*), $Br_{1.2}$ (*Mesure quadratique*), $Br_{2.1}$ (*GUEUC*), $Br_{2.2}$ (*Mesure Omni-monotone*), $Br_{2.3}$ (*Mesure Opti-monotone*), Br_3 (*Mesure anti-monotone*), voir *page 32* pour les définitions de ces critères.

Par la confrontation de nos travaux, nous cherchons à identifier si les mesures communes, qui sont à la fois stables selon notre étude empirique et groupées ensemble selon les études menées dans les *chapitres 3* et *4*, possèdent aussi un même comportement selon [Bra11]. Nous étudions chacun des groupes stables (*section 5.5, page 157*) de mesures séparément et nous nous apercevons que :

– G_{St1} : il regroupe les mesures *Nouveauté* et *Piatetsky-shapiro*, deux mesures optimonotones, qui ne sont pas planes et ne vérifient pas la propriété GUEUC. *Nouveauté* semble être d'une part plus robuste que *Piatetsky-shapiro* puisqu'elle est quadratique et possède une bonne propriété d'anti-monotonie dans le cas de règles de classes. Ainsi, ces deux mesures évaluent 4 propriétés parmi 6 de la même manière ;
– G_{St2} : ce groupe est divisé en 2 sous-classes selon les résultats de la classification décrite dans le *chapitre 3, page 89*. Nous avons :

  1. *Risque relatif, Facteur de certitude* : ces deux mesures vérifient les 6 critères de la même manière, elles sont quadratiques et opti-monotones ;
  2. *Taux d'exemples, Sebag, Ganascia, Laplace* : sont des mesures opti, omni-monotones, et non quadratiques. Nous nous apercevons aussi que les trois mesures *Taux d'exemples, Sebag* et *Ganascia* évaluent toutes les propriétés de la même

manière. Elles possèdent des lignes de niveau qui forment des plans et vérifient la propriété GUEUC.

- G_{St4} : regroupe les mesures *Gain informationnel* et *Intérêt*. Ces deux mesures possèdent aussi un même comportement selon [Bra11] puisqu'elles évaluent les 6 propriétés pareillement. Elles sont quadratiques, opti-monotones et possèdent la propriété GUEUC. Elles possèdent donc selon l'auteur des propriétés opérationnelles fortes ;
- G_{St5} : ce groupe stable comprend les mesures *Czekanowski-dice, Jaccard* et *Kulczynski* possèdant 5 propriétés communes parmi 6. Toutes ces mesures sont planes, opti-monotones et possèdent la propriété d'anti-monotonie pour les règles de classes. *Kulczynski* est la seule qui possède la propriété GUEUC ;
- G_{St8} : contient les mesures *Conviction* et *Facteur bayésien* qui possèdent aussi un comportement similaire selon l'auteur en vérifiant toutes les propriétés de la même manière. Elles sont donc quadratiques, possédant les propriétés de GUEUC et d'optimonotonie. Dans Lenca et al. [LMVL08], ces deux mesures se trouvent en tête suivant l'un des scénarii qu'il propose.

Après avoir effectué une étude comparative avec les résultats des travaux existants dans la littérature, et réalisés sur la classification des mesures selon un cadre empirique, nous reprenons maintenant la *table 1.6, page 44*, afin de positionner notre travail par rapport à l'ensemble des études empiriques existants.

La *table 5.18* résume les différentes travaux effectués sur les mesures d'intérêt objectives selon un point de vue analyse de données et met en valeur notre contribution.

5.8 Conclusion

Dans ce chapitre, nous présentons une étude comparative expérimentale du comportement de 60 mesures sur 6 jeux de données de nature différente (*dense/éparse, réel/synthétique*). Cette approche est considérée comme étant une étude complémentaire aux travaux que nous avons présentés dans les *chapitres 3 et 4*, puisque nous cherchons à comprendre le comportement des mesures d'intérêt expérimentalement et selon différentes bases de données. Ainsi, nous avons comparé le comportement de 60 mesures selon 6 jeux de données distincts, suite à l'évaluation d'une connaissance émise sous forme de règles d'association. En effet, il s'agit d'extraire les *N meilleures règles* par chacune des mesures et sur chaque jeu de données et de les comparer par la suite. Cette approche a débouché sur la construction de matrices de similarité entre mesures pour chaque jeu de données à partir desquelles nous avons cherché à interpréter le comportement des mesures par l'analyse de celles-ci. Ainsi, 3 catégories de me-

sures sont identifiées : (i) mesures au comportement similaire , (ii) mesures au comportement différent et (iii) mesures au comportement indéterminé et 8 groupes de mesures stables sont retenus. Ces groupes stables permettent de faciliter le travail de l'utilisateur face au problème de sélection de mesures objectives. Ce dernier n'a plus qu'à choisir aléatoirement une mesure de chaque groupe stable, puisque les mesures d'un même groupe stable sont censées proposer les mêmes *N* meilleures règles.

En outre, nous avons confronté les résultats empiriques obtenus par cette étude avec les 7 classes de mesures retenues dans le *chapitre 3* et nous avons remarqué que nous sommes incapable de valider toutes les classes : une classe est vérifiée, 2 classes ne le sont pas et 4 classes sont partiellement vérifiées. Parmi ces 4 classes, nous rappelons que dans la *section 3.4, page 90*, certaines ont été divisées en des sous-groupes afin de faciliter leur interprétation. Ainsi, l'étude théorique que nous avons réalisée sur les mesures reste toujours approximative, qui dépend des méthodes de classification utilisées et du nombre de propriétés étudiées.

Aide au choix des mesures :

Afin de faciliter la tâche de l'utilisateur, à choisir la/les mesure(s) la/les mieux appropriée(s) à ses besoins, nous suggérons de considérer les deux résultats de la classification que nous avons obtenus (*selon une approche formelle et empirique*). Comme nous avons discerné 7 classes de mesures par l'étude théorique et 8 sous-groupes stables de mesures selon l'étude empirique, l'utilisateur se retrouve en confusion quant au choix de la(les) bonne(s) mesure(s).

Toutefois, nous tenons tout d'abord à mentionner que nos résultats restent approximatives. D'un point de vue formel, les méthodes de classification utilisées ainsi que le nombre de propriétés étudiées jouent un rôle important dans l'étape de classification des mesures. Par exemple, l'étude de propriétés additionnelles peut engendrer une classification différente pour certaines mesures. D'un point de vue empirique, nous avons remarqué qu'en comparant nos résultats avec des classifications existantes, il y a des dissimilarités dans nos résultats, qui peuvent-être engendrées par le choix des jeux de données analysées. Ainsi, puisque les mesures d'intérêt dépendent de la nature des données, nous pouvons dire que nos résultats empiriques restent préliminaires. Nous avons obtenus 3 catégories de mesures selon notre approche empirique et il se peut qu'en choisissant d'autres jeux de données, les mesures appartenant à la *catégorie 1* et *catégorie 3*, appartiendront à la *catégorie 2*.

Dans un article récent de Tew et al. [TGCTB13], les auteurs se sont concentrés sur l'analyse du comportement de classement de règles de 61 mesures d'intérêt. Pour ce faire, ils ont réalisé une série de tests sur des règles générées à partir de 110 jeux de données différents et ils ont catégorisé les mesures selon les classements de règles obtenus, en cherchant les

mesures fortement corrélées, faiblement corrélées, indépendantes, etc. Les auteurs concluent que la présence de groupes de mesures distincts confirme que la connaissance du domaine est essentielle lors de la sélection d'une mesure d'intérêt, qui soit appropriée aux attentes et aux objectifs de l'utilisateur.

Néanmoins, étant donnée l'étude approfondie que nous avons réalisée sur un nombre important de mesures, nous suggérons de tenir compte de nos 2 classifications à la fois. La *figure 5.16*, retenue des résultats de notre étude théorique, illustre des classes de mesures de taille réduite que celles de la *page 89*. En effet, pour chaque classe de mesures, et dans le cas où un groupe stable identifié dans l'étude empirique est présent, nous ne gardons qu'une seule mesure, choisie aléatoirement (G_{St1} *: Piatetsky-shapiro* $\in G_{p8}$*, Leverage* $\in C_4$*, Pearl,* G_{St2} *:* M_{GK}*, Risque relatif* $\in C_7$*, Laplace* $\in C_5$*,* G_{St3} *: Information mutuelle* $\in C_3$*,* G_{St4} *: Intérêt* $\in C_7$*,* G_{St5} *: Jaccard* $\in C_4$*,* G_{St6} *: Confiance causale* $\in C_4$*,* G_{St7} *: IP3E* $\in C_2$*, Rappel* $\in C_4$*,* G_{St8} *: Conviction* $\in C_7$) de chaque groupe stable. L'utilisateur pourra alors commencer ses tests par les mesures objectives, jugées représentatives de leur classe dans la *section 3.4, page 90*, et qui sont colorées en rouge. Certaines classes ont été divisées en des sous-classes lors de leur interprétation dans le *chapitre 3*. Pour ces classes, nous proposons les mesures qu'il serait souhaitable pour l'utilisateur de les examiner, celles qui sont colorées en bleu sont les deuxièmes mesures à considérer, puis celles en vert.

Bien que le nombre de mesures à sélectionner est réduit, cette réduction reste toujours limitée puisque la plupart des mesures se concentrent généralement sur des caractéristiques différentes et uniques des règles. L'intérêt reste essentiellement subjectif [Sah99].

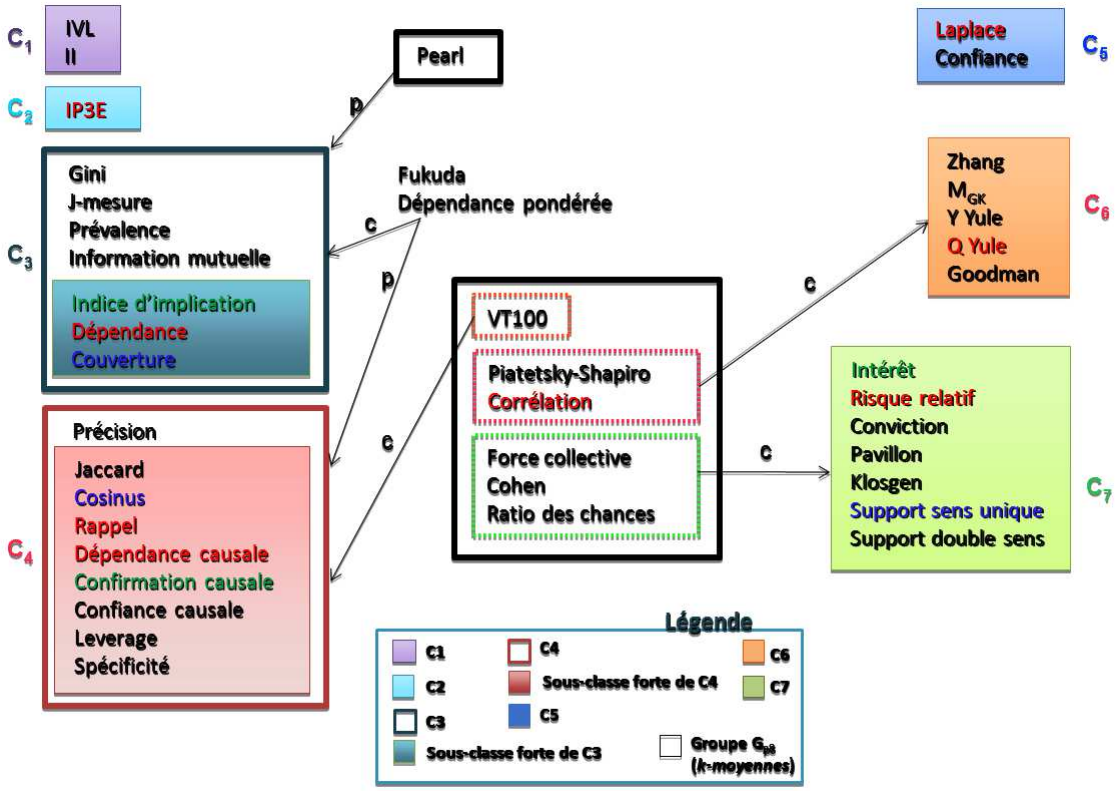

FIGURE *5.16: Mesures représentatives des classes avec lesquelles l'utilisateur peut commencer ses tests.*

Points clés

Positionnement :

– *Étude empirique du comportement d'une soixantaine de mesures d'intérêt selon des jeux de données de nature différentes.*

Contribution :

– *Classification des mesures d'intérêt d'un point de vue empirique* ;
– *Obtention de* 3 *catégories de mesures et identification de groupes de mesures ayant un comportement similaire*.

Publications :

– *D. Grissa, S. Guillaume, S. Ben Yahia, E. Mephu Nguifo (2011). Étude expérimentale des mesures d'intérêt pour l'extraction des connaissances*. Dans l'Atelier Extraction et Contextualisaton des Connaissances, AFIA'2011, Chambéry-FRANCE.
– *D. Grissa (2013). Étude comportementale de mesures d'intérêt de règles d'association*. Dans $11^{\text{èmes}}$ Rencontres des Jeunes Chercheurs en Intelligence Artificielle, RJCIA'13, Lille-France.

	Mesure	Cat1	Cat2	Cat3	VCI
Classe 1	IVL			×	NV
	II		×		
Classe 2	IEE	×			V
	IPEE	×			
	IP3E	×			
	IIER	×			
Classe 3	Indice d'implication		×		NV
	Gini		×		
	J-mesure			×	
	Dépendance			×	
	Prévalence		×		
	Couverture		×		
	Information mutuelle	×			
	Variation support (p)	×			
	Pearl (p)			×	
	Fukuda (c)			×	
	Dépendance pondérée (c)			×	
Classe 4	Précision		×		VP
	Jaccard	×			
	Support	×			
	Cosinus			×	
	Rappel	×			
	Dépendance causale			×	
	Confirmation causale			×	
	Confiance causale	×			
	Confiance confirmée-causale	×			
	Fiabilité négative	×			
	Leverage			×	
	Spécificité			×	
	Czekanowski-dice	×			
	Kulczynski	×			
Classe 5	Sebag	×			VP
	Moindre contradiction			×	
	Confirmation descriptive		×		
	Taux d'exemples	×			
	Ganascia	×			
	Laplace	×			
	Confiance			×	
Classe 6	Zhang			×	VP
	M_{GK}	×			
	Y de Yule			×	
	Q de Yule			×	
	Goodman			×	

TABLE *5.14: Comportement des mesures selon l'étude empirique (Partie 1).*

	Mesure	Cat1	Cat2	Cat3	VCl
Classe 7	Intérêt	×			VP
	Gain informationnel	×			
	Risque relatif	×			
	Facteur bayésien	×			
	Conviction	×			
	Facteur de certitude	×			
	Pavillon			×	
	Klosgen			×	
	Support sens unique			×	
	Support double sens		×		
Classe Gp_8	VT100			×	VP
	Piatetsky-shapiro	×			
	Correlation			×	
	Nouveauté	×			
	Force collective			×	
	Cohen			×	
	Ratio des chances			×	

TABLE *5.15: Comportement des mesures selon l'étude empirique (Partie 2).*

	Sim	Conf	Int	Lev	Conv	FN	Nov	Rap	Lap	MC	Seb	PS	Prec	Spec	RR	TEC	RC	Pearl	Prev	Cor	Coh	FCol	Pav	Gan	CfmD	Cos	Cov	CzD	FB	FC	GI	Klos	Kulz	SSU	SDS	CConf	CCC	CfmC	Dep	PCD	DP	Fuk	Gini	IndImp	Jac	Jmes	Mgk	QY	YY	VS	Zhang	IM	II	IPEE	IP3E	IIER	IIE	IVL	VT100	Good	Sup
Conf	I_S	**1,00**																																																											
	σ	**0,00**																																																											
Int	I_S	0,00	**1,00**																																																									
	σ	0,00	**0,00**																																																									
Lev	I_S	0,00	0,43	**1,00**																																																							
	σ	0,00	0,34	**0,00**																																																							
Conv	I_S	0,00	0,18	0,31	**1,00**																																																				
	σ	0,01	0,40	0,46	**0,00**																																																				
FN	I_S	0,00	0,18	0,21	0,72	**1,00**																																																		
	σ	0,01	0,40	0,40	0,36	**0,00**																																																		
Nov	I_S	0,00	0,43	1,00	0,31	0,21	**1,00**																																																	
	σ	0,00	0,34	0,00	0,46	0,40	**0,00**																																																	
Rap	I_S	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**																																															
	σ	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**																																															
Lap	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**																																													
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	**0,00**																																													
MC	I_S	0,37	0,00	0,00	0,00	0,00	0,00	0,00	0,65	**1,00**																																											
	σ	0,46	0,00	0,00	0,01	0,01	0,00	0,00	0,36	**0,00**																																											
Seb	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,65	**1,00**																																										
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	**0,00**																																										
PS	I_S	0,00	0,43	0,98	0,31	0,21	0,98	0,00	0,00	0,00	0,00	**1,00**																																									
	σ	0,00	0,34	0,04	0,46	0,40	0,04	0,00	0,00	0,00	0,00	**0,00**																																									
Prec	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**																																								
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**																																								
Spec	I_S	0,00	0,19	0,31	0,34	0,21	0,31	0,00	0,00	0,00	0,00	0,31	0,01	**1,00**																																						
	σ	0,00	0,40	0,46	0,49	0,39	0,46	0,00	0,00	0,00	0,00	0,46	0,02	**0,00**																																						
RR	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,65	1,00	0,00	0,00	0,00	**1,00**																																					
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,00	0,00	0,00	0,00	**0,00**																																					
TEC	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,65	1,00	0,00	0,00	0,00	1,00	**1,00**																																				
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,00	0,00	0,00	0,00	0,00	**0,00**																																				
RC	I_S	0,00	0,17	0,30	0,30	0,19	0,30	0,00	0,00	0,00	0,00	0,30	0,00	0,30	0,00	0,00	**1,00**																																		
	σ	0,00	0,41	0,44	0,45	0,40	0,44	0,00	0,00	0,00	0,00	0,44	0,00	0,45	0,00	0,00	**0,00**																																		
Pearl	I_S	0,00	0,36	0,98	0,31	0,21	0,98	0,00	0,00	0,00	0,00	0,97	0,00	0,31	0,00	0,00	0,30	**1,00**																																	
	σ	0,00	0,39	0,04	0,46	0,40	0,04	0,00	0,00	0,00	0,00	0,05	0,00	0,46	0,00	0,00	0,44	**0,00**																																	
Prev	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**																																
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	**0,00**																																
Cor	I_S	0,00	0,00	0,01	0,00	0,00	0,01	0,00	0,01	0,01	0,01	0,01	0,00	0,00	0,01	0,01	0,23	0,01	0,00	**1,00**																															
	σ	0,00	0,00	0,03	0,01	0,01	0,03	0,00	0,01	0,01	0,01	0,03	0,01	0,00	0,01	0,01	0,35	0,03	0,00	**0,00**																															
Coh	I_S	0,00	0,02	0,02	0,00	0,00	0,02	0,00	0,01	0,01	0,01	0,02	0,00	0,00	0,01	0,01	0,20	0,02	0,00	0,60	**1,00**																														
	σ	0,00	0,02	0,03	0,01	0,01	0,03	0,00	0,01	0,01	0,01	0,03	0,01	0,00	0,01	0,01	0,31	0,03	0,00	0,44	**0,00**																														
FCol	I_S	0,00	0,18	0,31	0,30	0,19	0,31	0,00	0,00	0,00	0,00	0,31	0,00	0,29	0,00	0,00	0,53	0,31	0,00	0,19	0,29	**1,00**																													
	σ	0,00	0,40	0,44	0,45	0,40	0,44	0,00	0,00	0,00	0,00	0,44	0,00	0,45	0,00	0,00	0,45	0,44	0,00	0,30	0,41	**0,00**																													
Pav	I_S	0,00	0,19	0,33	0,38	0,23	0,33	0,00	0,00	0,00	0,00	0,33	0,00	0,33	0,00	0,00	0,30	0,33	0,00	0,01	0,00	0,30	**1,00**																												
	σ	0,00	0,40	0,45	0,48	0,38	0,45	0,00	0,00	0,00	0,00	0,45	0,00	0,49	0,00	0,00	0,45	0,45	0,00	0,02	0,01	0,44	**0,00**																												
Gan	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,65	1,00	0,00	0,00	0,00	1,00	1,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	**1,00**																											
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	**0,00**																											
CfmD	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	**1,00**																										
	σ	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	**0,00**																										
Cos	I_S	0,00	0,01	0,00	0,00	0,00	0,00	0,15	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,33	0,01	0,00	0,01	0,01	**1,00**																									
	σ	0,00	0,02	0,00	0,00	0,00	0,00	0,17	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,00	0,50	0,52	0,02	0,00	0,01	0,02	**0,00**																									
Cov	I_S	0,03	0,00	0,00	0,00	0,00	0,00	0,01	0,03	0,04	0,03	0,00	0,00	0,00	0,03	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,00	0,00	**1,00**																								
	σ	0,07	0,00	0,00	0,00	0,00	0,00	0,01	0,07	0,10	0,07	0,00	0,00	0,00	0,07	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,00	0,01	**0,00**																								
CzD	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,28	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,36	0,01	0,00	0,01	0,02	0,69	0,00	**1,00**																							
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,32	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,00	0,50	0,50	0,02	0,00	0,01	0,02	0,30	0,01	**0,00**																							
FB	I_S	0,00	0,18	0,31	0,95	0,71	0,31	0,00	0,00	0,00	0,00	0,31	0,00	0,34	0,00	0,00	0,30	0,31	0,00	0,00	0,00	0,30	0,38	0,00	0,00	0,00	0,00	0,00	**1,00**																						
	σ	0,00	0,40	0,46	0,06	0,35	0,46	0,00	0,00	0,00	0,00	0,46	0,00	0,48	0,00	0,00	0,45	0,46	0,00	0,01	0,01	0,45	0,48	0,00	0,00	0,00	0,00	0,00	**0,00**																						
FC	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,65	1,00	0,00	0,00	0,00	1,00	1,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	1,00	0,00	0,01	0,03	0,01	0,00	**1,00**																					
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,00	0,01	0,07	0,01	0,00	**0,00**																					
GI	I_S	0,00	1,00	0,43	0,18	0,18	0,43	0,00	0,00	0,00	0,00	0,43	0,00	0,19	0,00	0,00	0,17	0,36	0,00	0,00	0,02	0,18	0,19	0,00	0,00	0,01	0,00	0,00	0,18	0,00	**1,00**																				
	σ	0,00	0,00	0,34	0,40	0,40	0,34	0,00	0,00	0,00	0,00	0,34	0,00	0,40	0,00	0,00	0,41	0,39	0,00	0,00	0,02	0,40	0,40	0,00	0,00	0,02	0,00	0,00	0,40	0,00	**0,00**																				
Klos	I_S	0,00	0,19	0,34	0,38	0,23	0,34	0,00	0,00	0,00	0,00	0,34	0,00	0,33	0,00	0,00	0,29	0,34	0,00	0,00	0,00	0,30	0,61	0,00	0,00	0,00	0,00	0,00	0,38	0,00	0,19	**1,00**																			
	σ	0,00	0,40	0,44	0,48	0,38	0,44	0,00	0,00	0,00	0,00	0,44	0,00	0,49	0,00	0,00	0,45	0,44	0,00	0,00	0,00	0,44	0,33	0,00	0,00	0,00	0,00	0,00	0,48	0,00	0,40	**0,00**																			
Kulz	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,28	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,36	0,01	0,00	0,01	0,02	0,69	0,00	1,00	0,00	0,01	0,00	0,00	**1,00**																		
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,32	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,00	0,50	0,50	0,02	0,00	0,01	0,02	0,30	0,01	0,00	0,00	0,01	0,00	0,00	**0,00**																		
SSU	I_S	0,18	0,00	0,00	0,00	0,00	0,00	0,00	0,18	0,19	0,18	0,00	0,00	0,00	0,18	0,18	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,18	0,00	0,00	0,00	0,00	0,00	0,18	0,00	0,00	0,00	**1,00**																	
	σ	0,40	0,00	0,00	0,01	0,01	0,00	0,00	0,40	0,40	0,40	0,00	0,00	0,00	0,40	0,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,40	0,00	0,00	0,00	0,00	0,00	0,40	0,00	0,00	0,00	**0,00**																	
SDS	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	**1,00**																
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	**0,00**																

TABLE *5.16: Matrice complète présentant les valeurs de l'indice de similarité I_S et de l'écart-type σ obtenus par l'étude empirique en comparant les couples de mesures d'intérêt selon 6 jeux de données (1). Les abréviations des différentes mesures se trouve dans la table C.1, page 210.*

	Sim	Conf	Int	Lev	Conv	FN	Nov	Rap	Lap	MC	Seb	PS	Prec	Spec	RR	TEC	RC	Pearl	Prev	Cor	Coh	FCol	Pav	Gan	CfmD	Cos	Cov	CzD	FB	FC	GI	Klos	Kulz	SSU	SDS
CConf	I_S	0,00	0,18	0,21	0,72	1,00	0,21	0,00	0,00	0,00	0,00	0,21	0,00	0,21	0,00	0,00	0,19	0,21	0,00	0,00	0,00	0,19	0,23	0,00	0,00	0,00	0,00	0,00	0,71	0,00	0,18	0,23	0,00	0,00	0,00
	σ	0,01	0,40	0,40	0,36	0,00	0,40	0,00	0,01	0,01	0,01	0,40	0,00	0,39	0,01	0,01	0,40	0,40	0,00	0,01	0,01	0,40	0,38	0,01	0,00	0,00	0,00	0,00	0,35	0,01	0,40	0,38	0,00	0,01	0,00
CCC	I_S	0,00	0,18	0,21	0,72	1,00	0,21	0,00	0,00	0,00	0,00	0,21	0,00	0,21	0,00	0,00	0,19	0,21	0,00	0,00	0,00	0,19	0,23	0,00	0,00	0,00	0,00	0,00	0,71	0,00	0,18	0,23	0,00	0,00	0,00
	σ	0,01	0,40	0,40	0,36	0,00	0,40	0,00	0,01	0,01	0,01	0,40	0,00	0,39	0,01	0,01	0,40	0,40	0,00	0,01	0,01	0,40	0,38	0,01	0,00	0,00	0,00	0,00	0,35	0,01	0,40	0,38	0,00	0,01	0,00
CfmC	I_S	0,00	0,00	0,00	0,01	0,01	0,00	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,34	0,36	0,00	0,00	0,01	0,04	0,37	0,00	0,36	0,01	0,01	0,00	0,03	0,36	0,00	0,00
	σ	0,00	0,00	0,00	0,02	0,02	0,00	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,02	0,51	0,50	0,00	0,01	0,01	0,04	0,49	0,00	0,50	0,02	0,01	0,00	0,07	0,50	0,00	0,00
Dep	I_S	0,00	0,18	0,33	0,38	0,23	0,33	0,00	0,00	0,00	0,00	0,33	0,00	0,33	0,00	0,00	0,30	0,33	0,00	0,01	0,00	0,30	0,88	0,00	0,00	0,00	0,00	0,00	0,38	0,00	0,18	0,57	0,00	0,00	0,00
	σ	0,00	0,40	0,45	0,48	0,38	0,45	0,00	0,00	0,00	0,00	0,45	0,00	0,49	0,00	0,00	0,45	0,45	0,00	0,02	0,01	0,44	0,28	0,00	0,00	0,00	0,00	0,00	0,48	0,00	0,40	0,39	0,00	0,00	0,00
PCD	I_S	0,01	0,00	0,00	0,10	0,10	0,00	0,00	0,11	0,11	0,11	0,00	0,00	0,00	0,11	0,11	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,11	0,00	0,01	0,00	0,01	0,10	0,11	0,00	0,01	0,01	0,01	0,00
	σ	0,02	0,00	0,00	0,23	0,23	0,00	0,00	0,17	0,17	0,17	0,00	0,00	0,00	0,17	0,17	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,17	0,00	0,01	0,00	0,01	0,23	0,17	0,00	0,02	0,01	0,01	0,00
DP	I_S	0,00	0,27	0,44	0,32	0,21	0,44	0,00	0,00	0,00	0,00	0,44	0,00	0,32	0,00	0,00	0,29	0,44	0,01	0,00	0,00	0,30	0,32	0,00	0,00	0,00	0,00	0,01	0,32	0,00	0,27	0,32	0,01	0,00	0,00
	σ	0,00	0,38	0,44	0,46	0,39	0,44	0,00	0,00	0,00	0,00	0,44	0,00	0,45	0,00	0,00	0,45	0,44	0,03	0,01	0,00	0,44	0,46	0,00	0,00	0,00	0,00	0,01	0,46	0,00	0,38	0,45	0,01	0,00	0,00
Fuk	I_S	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,01	0,05	0,01	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,00	0,00
	σ	0,02	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,02	0,02	0,00	0,00	0,00	0,02	0,02	0,00	0,00	0,02	0,00	0,01	0,00	0,00	0,02	0,05	0,01	0,00	0,01	0,00	0,02	0,00	0,00	0,01	0,00	0,00
Gini	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
IndImp	I_S	0,00	0,00	0,00	0,04	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	σ	0,00	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Jac	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,28	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,36	0,01	0,00	0,01	0,02	0,69	0,00	1,00	0,00	0,01	0,00	0,00	1,00	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,32	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,00	0,50	0,50	0,02	0,00	0,01	0,02	0,30	0,01	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00
Jmes	I_S	0,00	0,18	0,31	0,50	0,38	0,31	0,00	0,00	0,00	0,00	0,31	0,00	0,33	0,00	0,00	0,29	0,31	0,00	0,00	0,00	0,29	0,39	0,00	0,00	0,00	0,00	0,00	0,54	0,00	0,18	0,39	0,00	0,00	0,00
	σ	0,00	0,40	0,46	0,55	0,45	0,46	0,00	0,00	0,00	0,00	0,46	0,00	0,49	0,00	0,00	0,45	0,46	0,00	0,00	0,00	0,45	0,48	0,00	0,00	0,00	0,00	0,00	0,51	0,00	0,40	0,48	0,00	0,00	0,00
Mgk	I_S	0,58	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,65	1,00	0,00	0,00	0,00	1,00	1,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	1,00	0,00	0,01	0,03	0,01	0,00	1,00	0,00	0,00	0,01	0,18	0,00
	σ	0,43	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,00	0,01	0,07	0,01	0,00	0,00	0,00	0,00	0,01	0,40	0,00
YQ	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,02	0,02	0,00	0,00	0,00	0,00	0,05	0,00	0,05	0,00	0,00	0,00	0,00	0,05	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,06	0,00	0,01	0,06	0,06	0,00	0,00	0,01	0,00	0,12	0,01	0,12	0,00	0,01	0,00	0,00	0,12	0,00	0,00
YY	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,25	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,10	0,01	0,18	0,00	0,00	0,00	0,00	0,18	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,31	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,02	0,11	0,01	0,23	0,00	0,00	0,00	0,00	0,23	0,00	0,00
VS	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,39	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,13	0,01	0,27	0,00	0,00	0,00	0,00	0,27	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,43	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,15	0,01	0,31	0,00	0,00	0,00	0,00	0,31	0,00	0,00
Zhang	I_S	0,42	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,59	0,83	0,00	0,00	0,00	0,83	0,83	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,83	0,00	0,01	0,03	0,01	0,00	0,83	0,00	0,00	0,01	0,17	0,00
	σ	0,44	0,00	0,00	0,01	0,01	0,00	0,00	0,41	0,43	0,41	0,00	0,00	0,00	0,41	0,41	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,41	0,00	0,01	0,07	0,01	0,00	0,41	0,00	0,00	0,01	0,41	0,00
IM	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,39	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,13	0,01	0,27	0,00	0,00	0,00	0,00	0,27	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,43	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,15	0,01	0,31	0,00	0,00	0,00	0,00	0,31	0,00	0,00
II	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,00	0,00
	σ	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,01	0,01	0,00
IPEE	I_S	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,15	0,01	0,28	0,00	0,00	0,00	0,00	0,28	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,17	0,01	0,32	0,00	0,00	0,00	0,00	0,32	0,00	0,00
IP3E	I_S	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,15	0,01	0,28	0,00	0,00	0,00	0,00	0,28	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,17	0,01	0,32	0,00	0,00	0,00	0,00	0,32	0,00	0,00
IIER	I_S	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,15	0,01	0,28	0,00	0,00	0,00	0,00	0,28	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,17	0,01	0,32	0,00	0,00	0,00	0,00	0,32	0,00	0,00
IIE	I_S	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,15	0,01	0,28	0,00	0,00	0,00	0,00	0,28	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,17	0,01	0,32	0,00	0,00	0,00	0,00	0,32	0,00	0,00
IVL	I_S	0,00	0,15	0,11	0,00	0,00	0,11	0,33	0,00	0,00	0,00	0,12	0,00	0,00	0,00	0,00	0,00	0,11	0,00	0,00	0,00	0,01	0,02	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,15	0,02	0,00	0,00	0,00
	σ	0,00	0,28	0,20	0,00	0,00	0,20	0,52	0,00	0,00	0,00	0,21	0,00	0,00	0,00	0,00	0,00	0,20	0,00	0,00	0,01	0,02	0,03	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,28	0,03	0,00	0,00	0,00
VT100	I_S	0,00	0,17	0,17	0,17	0,17	0,17	0,00	0,01	0,01	0,01	0,17	0,00	0,17	0,01	0,01	0,24	0,17	0,00	0,18	0,17	0,17	0,17	0,01	0,00	0,17	0,00	0,17	0,17	0,01	0,17	0,17	0,17	0,00	0,00
	σ	0,01	0,41	0,41	0,41	0,41	0,41	0,00	0,01	0,01	0,01	0,41	0,01	0,41	0,01	0,01	0,40	0,41	0,00	0,40	0,41	0,41	0,41	0,01	0,01	0,41	0,00	0,41	0,41	0,01	0,41	0,41	0,41	0,00	0,00
Good	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,00	0,00
	σ	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,02	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,00	0,00
Sup	I_S	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,15	0,01	0,28	0,00	0,00	0,00	0,00	0,28	0,00	0,00
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,17	0,01	0,32	0,00	0,00	0,00	0,00	0,32	0,00	0,00

	Sim	CConf	CCC	CfmC	Dep	PCD	DP	Fuk	Gini	IndImp	Jac	Jmes	Mgk	QY	YY	VS	Zhang	IM	II	IPEE	IP3E	IIER	IIE	IVL	VT100	Good	Sup
CConf	I_S	**1,00**																									
	σ	**0,00**																									
CCC	I_S	1,00	**1,00**																								
	σ	0,00	**0,00**																								
CfmC	I_S	0,01	0,01	**1,00**																							
	σ	0,02	0,02	**0,00**																							
Dep	I_S	0,23	0,23	0,00	**1,00**																						
	σ	0,38	0,38	0,01	**0,00**																						
PCD	I_S	0,10	0,10	0,01	0,00	**1,00**																					
	σ	0,23	0,23	0,01	0,00	**0,00**																					
DP	I_S	0,21	0,21	0,00	0,32	0,00	**1,00**																				
	σ	0,39	0,39	0,01	0,45	0,01	**0,00**																				
Fuk	I_S	0,00	0,00	0,05	0,00	0,01	0,00	**1,00**																			
	σ	0,00	0,00	0,06	0,00	0,02	0,00	**0,00**																			
Gini	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**																		
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**																		
IndImp	I_S	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**																	
	σ	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	**0,00**																	
Jac	I_S	0,00	0,00	0,36	0,00	0,01	0,00	0,01	0,00	0,00	**1,00**																
	σ	0,00	0,00	0,50	0,00	0,01	0,00	0,01	0,00	0,00	**0,00**																
Jmes	I_S	0,38	0,38	0,01	0,39	0,00	0,32	0,00	0,00	0,00	0,00	**1,00**															
	σ	0,45	0,45	0,02	0,48	0,00	0,46	0,00	0,00	0,00	0,00	**0,00**															
Mgk	I_S	0,00	0,00	0,01	0,00	0,11	0,00	0,01	0,00	0,00	0,01	0,00	**1,00**														
	σ	0,01	0,01	0,01	0,00	0,17	0,00	0,02	0,00	0,00	0,01	0,00	**0,00**														
YQ	I_S	0,00	0,00	0,03	0,00	0,00	0,00	0,01	0,00	0,00	0,05	0,00	0,00	**1,00**													
	σ	0,00	0,00	0,06	0,00	0,01	0,00	0,01	0,00	0,00	0,12	0,01	0,01	**0,00**													
YY	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,18	0,00	0,00	0,01	**1,00**												
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,23	0,00	0,00	0,01	**0,00**												
VS	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,27	0,00	0,00	0,01	0,67	**1,00**											
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,31	0,00	0,00	0,01	0,40	**0,00**											
Zhang	I_S	0,00	0,00	0,01	0,00	0,14	0,00	0,01	0,00	0,00	0,01	0,00	0,83	0,00	0,00	0,00	**1,00**										
	σ	0,01	0,01	0,01	0,00	0,17	0,00	0,02	0,00	0,00	0,01	0,00	0,41	0,01	0,00	0,00	**0,00**										
IM	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,27	0,00	0,00	0,01	0,67	1,00	0,00	**1,00**									
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,31	0,00	0,00	0,01	0,40	0,00	0,00	**0,00**									
II	I_S	0,00	0,00	0,01	0,00	0,02	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,00	**1,00**								
	σ	0,00	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,01	0,00	**0,00**								
IPEE	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,28	0,00	0,00	0,01	0,25	0,39	0,00	0,39	0,00	**1,00**							
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,32	0,00	0,00	0,01	0,31	0,43	0,00	0,43	0,00	**0,00**							
IP3E	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,28	0,00	0,00	0,01	0,25	0,39	0,00	0,39	0,00	1,00	**1,00**						
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,32	0,00	0,00	0,01	0,31	0,43	0,00	0,43	0,00	0,00	**0,00**						
IIER	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,28	0,00	0,00	0,01	0,25	0,39	0,00	0,39	0,00	1,00	1,00	**1,00**					
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,32	0,00	0,00	0,01	0,31	0,43	0,00	0,43	0,00	0,00	0,00	**0,00**					
IIE	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,28	0,00	0,00	0,01	0,25	0,39	0,00	0,39	0,00	1,00	1,00	1,00	**1,00**				
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,32	0,00	0,00	0,01	0,31	0,43	0,00	0,43	0,00	0,00	0,00	0,00	**0,00**				
IVL	I_S	0,00	0,00	0,00	0,02	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,33	0,33	0,33	0,33	**1,00**			
	σ	0,00	0,00	0,01	0,02	0,00	0,04	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,01	0,00	0,52	0,52	0,52	0,52	**0,00**			
VT100	I_S	0,17	0,17	0,17	0,17	0,01	0,17	0,00	0,00	0,00	0,17	0,17	0,01	0,00	0,00	0,00	0,01	0,00	0,02	0,00	0,00	0,00	0,00	0,00	**1,00**		
	σ	0,41	0,41	0,41	0,41	0,01	0,41	0,01	0,00	0,00	0,41	0,41	0,01	0,01	0,00	0,00	0,01	0,00	0,02	0,00	0,00	0,00	0,00	0,00	**0,00**		
Good	I_S	0,00	0,00	0,01	0,00	0,02	0,00	0,01	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,04	0,00	0,00	0,00	0,00	0,00	0,14	**1,00**	
	σ	0,00	0,00	0,01	0,00	0,01	0,00	0,02	0,00	0,00	0,01	0,00	0,01	0,01	0,00	0,00	0,01	0,00	0,04	0,00	0,00	0,00	0,00	0,00	0,24	**0,00**	
Sup	I_S	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,28	0,00	0,00	0,01	0,25	0,39	0,00	0,39	0,00	1,00	1,00	1,00	1,00	0,33	0,00	0,00	**1,00**
	σ	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,32	0,00	0,00	0,01	0,31	0,43	0,00	0,43	0,00	0,00	0,00	0,00	0,00	0,52	0,00	0,00	**0,00**

TABLE *5.17: Matrice complète présentant les valeurs de l'indice de similarité I_S et de l'écart-type σ obtenus par l'étude empirique en comparant les couples de mesures d'intérêt selon 6 jeux de données (2). Les abréviations des différentes mesures se trouve dans la table C.1, page 210.*

Auteurs	Nbre mesures	Nbre bases	Nature des bases	Type d'étude	Techniques utilisées	Résultats
Hyunh et al. [HGB05a]	34	1	Réelle/dense	Catégorisation des mesures d'intérêt	Plateforme d'expérimentation ARQAT + classification basée sur le graphe de corrélation positive	Identification de 11 groupes de mesures
Hyunh et al. [HGB05b]	35	1	Réelle/dense	catégorisation des mesures d'intérêt	approche d'analyse de données basée sur la distance entre les mesures + méthode de classification hiérarchique CAH + méthode de partitionnement k-médoide	Identification de 16 groupes de mesures
Hyunh et al. [HGB+07]	36	2	Dense/éparse	Catégorisation des mesures d'intérêt	Plateforme d'expérimentation ARQAT + classification basée sur le graphe de corrélation	Identification de 5 groupes de mesures
Vaillant et al. [VLL04] [LVML07]	20	10	Réelle	Catégorisation des mesures d'intérêt	Plateforme d'expérimentation HERBS + comparaison de préordres	Identification de 5 groupes de mesures
Plasse et al. [PKSL06]	5	1	Réelle/dense	Classement des règles + catégorisation des mesures	Représentation graphique originale basée sur des courbes de niveaux	Identification de deux groupes de mesures
Tan et al. [TKS02]	21	6	1 base synthétique + 5 bases réelles	Calcul de la similarité entres les paires de mesures + classement des tables de contingence + avis d'un expert	Calcul de la corrélation au moyen de l'indice de Pearson + ordonnancement au moyen d'un algorithme qu'ils proposent	Aucune mesure n'est meilleure que les autres pour tous les domaines d'application
Heravi et Zaiane [HZ10]	53	20	Réelle	Classification de règles associatives	Classifieurs associatifs	Il n'existe pas de mesure "*unique*" qui a un impact sur tous les ensembles de règles pour tous les jeux de données
E. Suzuki [Suz09b]	1	10	Réelle/artificielle	Génération de règles de classification	Méthode de découverte de règles CLARDER qu'ils ont proposé	Découverte de groupes de règles de classification
Hébert et Crémilleux [HC06]	17	1	Réelle/dense	Étude comparative des mesures d'intérêt + génération de règles de classification informatives	Proposition d'un environnement unificateur des mesures d'intérêt + d'une méthode de découverte de règles de classification	La majorité des mesures se comportent de la même manière
Hébert et Crémilleux [HC07]	17	1	Réelle/dense	Étude comparative des mesures d'intérêt + génération de règles d'association	Un nouvel environnement + nouveau algorithme sont proposés afin d'extraire les règles optimisant l'ensemble des mesures	Le choix d'une mesure d'intérêt n'est pas vraiment posé puisque toutes les mesures se comportent semblablement
Ohsaki et al. [OSK+04]	20	1	Réelle	Génération de règles d'association	Système d'extraction fondé sur un cadre typique de fouille de données, séries chronologiques + avis de l'expert	La combinaison de certaines mesures favorise l'interaction du système humain
Carvalho et al. [CFE05]	11	8	Réelle	Génération de règles de classification	Ordonnancement des règles + avis de l'expert	Il n'existe pas un "*gagnant*" clair parmi les mesures
Surana et al. [SKR10]	19	2	Réelle/éparse	Génération de règles d'association rares	Ordonnancement des règles + avis de l'expert	Aucune mesure unique n'est appropriée pour extraire les règles d'association rares pour tous les jeux de données
Notre contribution	60	6	Dense/éparse Réelle/artificielle	Classification des mesures d'intérêt	Plateforme d'expérimentation Weka + algorithme d'extraction Apriori	Identification de 3 catégories de mesures, où 8 sous-groupes stables au comportement similaire sont obtenus

TABLE *5.18: Tableau de synthèse positionnant notre contribution par rapport à l'ensemble des travaux réalisés sur les mesures d'intérêt selon une approche empirique.*

Conclusion générale et perspectives

Les travaux menés dans cette thèse ont porté essentiellement sur les mesures d'intérêt d'extraction de règles d'association. Ces mesures d'intérêt ont été proposées afin d'évaluer les règles et de proposer des connaissances intéressantes à l'utilisateur. Néanmoins, plusieurs mesures existent dans la littérature posant par conséquent un problème de sélection à l'utilisateur expert. En outre, vu le nombre prohibitif de règles qui peuvent être générées par les algorithmes d'extraction, il est indispensable de choisir la/les mesure(s) la/les mieux adaptée(s) aux besoins l'utilisateur. Ainsi et afin de l'aider dans sa tâche, nous avons répondu à cette problématique par la proposition d'une étude approfondie d'un nombre important de mesures d'intérêt selon deux points de vue : formel et empirique.

Les apports de notre approche

Les apports de cette thèse peuvent se résumer autour des points suivants :

- la mise en oeuvre d'un cadre formel pour l'étude des mesures d'intérêt existantes dans la littérature. Une soixantaine de mesures et une vingtaine de propriétés formelles ont été recensées. Ces propriétés de mesures ont été formalisées et analysées pour l'évaluation des mesures.
- la proposition d'une classification des mesures d'intérêt, basée sur des techniques à la fois sans et avec recouvrement :
  1. La première classification utilisant des techniques sans recouvrement (*CAH et k-moyennes*) a permis l'identification de 7 classes de mesures disjointes ;
  2. La deuxième classification a été réalisée au moyen de l'AFB, afin d'obtenir des groupes de mesures recouvrants ;
- la proposition de mesures représentatives pour chacune des classes de mesures obtenues par la classification disjointe. Ainsi, un nombre restreint de mesures sera proposé à l'utilisateur pour commencer ses tests ;
- l'identification de groupes de mesures stables par la comparaison de deux résultats de la classification non supervisée obtenue ;
- la mise en oeuvre d'une étude empirique sur le comportement des mesures d'intérêt en s'appuyant sur des jeux de données de type et nature différents. Cette étude empirique a permis de discerner 3 catégories de mesures : la première catégorie représente

les mesures au comportement similaire pour tous les jeux de données étudiés. Nous avons mis en exergue 8 sous-groupes de mesures stables. La deuxième catégorie représente les mesures au comportement différent, et la troisième catégorie correspond aux mesures dont le comportement est indéterminé.

Cadre formel

La première contribution s'est intéressée à l'établissement d'un cadre formel consistant en une étude approfondie du comportement des mesures d'intérêt d'extraction de règles d'association. Cette étude a porté sur les points suivants :

Évaluation des mesures d'intérêt Tout d'abord, nous avons exploré la littérature pour identifier l'ensemble de mesures et de propriétés existantes. Cette exploration a permis l'identification de 69 mesures d'intérêt et 22 propriétés formelles. Toutefois, parmi les 69 mesures recensées, nous avons détecté des groupes de mesures identiques, ayant une même définition mais des noms différents. Ainsi et pour éviter la redondance, nous avons gardé une seule mesure par groupe, diminuant par conséquent le nombre de mesures étudiées à 61 mesures.

Pour les propriétés de mesures, nous les avons toutes formalisées, à l'exception de 3 propriétés qui ont été dures à interpréter. Nous avons par la suite évalué les mesures d'intérêt selon les propriétés que nous avons formalisées, une évaluation qui a permis la construction d'une matrice de 61 mesures $\times$ 19 propriétés. Cette matrice d'évaluation était le point de départ pour une catégorisation de mesures.

Classification des mesures d'intérêt Afin d'aider l'utilisateur dans le choix de la mesure la mieux appropriée à ses besoins, nous avons effectué une classification des mesures d'intérêt recensées dans la littérature. Pour ce faire, nous avons appliqué sur la matrice d'évaluation des mesures selon les propriétés, différentes techniques de classification non supervisée. Nous avons choisi d'appliquer dans un premier temps des méthodes de classification sans recouvrement, qui sont la classification ascendante hiérarchique et la méthode des *k-moyennes*. Un consensus sur la classification a dévoilé 7 classes de mesures disjointes où chaque classe comprend un ensemble de mesures ayant des propriétés communes.

Cette première classification a été suivie d'une deuxième qui s'est intéressée au recouvrement des groupes de mesures. Cette nouvelle caractéristique a permis d'identifier des informations cachées dans les données, telles que les mesures qui peuvent appartenir à plus qu'une seule classe parce qu'elles partagent des propriétés avec chacune des classes. Cette

deuxième classification est obtenue par l'application de l'analyse factorielle booléenne sur la matrice d'évaluation des mesures. 38 facteurs ont été identifiés. Nous nous sommes intéressés aux 10 premiers facteurs uniquement puisqu'ils couvrent la totalité de la matrice. Nous avons aussi proposé une interprétation de ces derniers puisqu'ils représentent en effet nos groupes de mesures.

Mesure représentative Pour faciliter le travail de l'utilisateur lors du choix de mesures, nous avons proposé dans un premier temps des groupes de mesures partageant des propriétés communes. L'utilisateur peut choisir n'importe quelle mesure de chaque groupe pour effectuer ses tests. Néanmoins, cette étape de sélection de mesures a été encore améliorée par la proposition de mesures représentatives pour chaque groupe. Pour ce faire, nous avons cherché le "centre de gravité" de cette classe, et nous avons sélectionné la/les mesure(s) la/les plus proche(s) du centre comme mesure(s) représentante(s) de la classe.

Étude comparative Une autre contribution a consisté en une étude comparative des différents travaux réalisés sur la classification des mesures d'intérêt d'extraction de règles d'association. Une première étude a comparé les résultats de la classification disjointe obtenue avec des travaux existant dans la littérature. Cette comparaison a permis la détection de similarités au niveau du regroupement de plusieurs mesures communes.

Une deuxième étude a comparé les 7 classes de mesures obtenues par le consensus sur la classification entre la méthode CAH et celle des *k-moyennes*, avec les classes nommées fortes qui ont été détectées par la méthode des *k-moyennes* et finalement avec les groupes de mesures du diagramme de Venn discernés par la méthode d'analyse factorielle booléenne. Cette comparaison a révélé 7 groupes de mesures stables. Ces groupes stables déterminent les mesures qui sont toujours regroupées ensemble quelque soit la méthode de classification appliquée.

Cadre empirique

Ayant effectué une étude approfondie sur une soixantaine de mesures d'intérêt selon une approche formelle, nous avons effectué une étude complémentaire qui a consisté en l'analyse du comportement des mesures d'intérêt selon une approche empirique. Cette étude a été réalisée afin de valider les groupes de mesures d'intérêt obtenus par les méthodes de classification non supervisée, où les mesures appartenant à un même groupe devaient extraire les même *N* meilleures règles d'association.

Cette approche d'analyse de données a porté sur des bases de données spécifiques ayant des caractéristiques variées (*de nature et de type différents*) : 4 bases qui sont à la fois réelles et denses et 2 bases synthétiques et éparses. Notre approche a consisté à évaluer des connaissances produites sous forme de règles d'association. Elle s'est concentrée sur le calcul de la similarité entre les *N* meilleures règles extraites par chaque couple de mesures, où le taux de similarité a représenté l'union et l'intersection des *N* règles les mieux classées par les mesures. Cette étude empirique du comportement d'une soixantaine de mesures a dévoilé 3 catégories de mesures :

- des mesures au comportement semblable ;
- des mesures au comportement différent ;
- des mesures dont le comportement est indéterminé, en fonction des données.

Nous nous sommes intéressés essentiellement aux mesures de la première catégorie.

Groupes de mesures stables Nous avons étudié une notion importante, celle de groupes stables de mesures d'intérêt. Ces groupes sont généralement indépendants de la nature des données et de la sélection des règles d'association, i.e., les mesures d'un même groupe se comportent toujours semblablement quelque soit la nature des données. Par l'étude du comportement d'une soixantaine de mesures, nous avons pu identifier 8 sous-groupes stables de mesures auxquels nous avons proposé une interprétation.

L'interprétation de ces sous-groupes stables de mesures a révélé des relations intéressantes entre certaines mesures d'intérêt.

Étude comparative Les groupes de mesures stables au comportement similaire, ont été confrontés dans un premier temps aux groupes de mesures disjoints, obtenus par l'étude formelle. La comparaison de ces deux résultats a été effectuée afin de valider la classification formelle. Cette confrontation a montré que uniquement des sous-groupes de mesures d'intérêt de la classification formelle ont été validés.

Une deuxième étude comparative a été réalisée entre les 8 sous-groupes stables de mesures avec les résultats des travaux de la littérature. Cette comparaison a dévoilé des similarités dans le regroupement de certaines mesures communes.

Perspectives

Le présent travail soulève quelques voies pour des travaux futurs au niveau théorique et pratique, pouvant améliorer les résultats obtenus dans cette thèse :

1. Il serait intéressant d'envisager des propriétés complémentaires pour l'étude du comportement des mesures, comme la notion de robustesse. Cette notion consiste à observer les variations des mesures au moment où des modifications sont effectuées au niveau des données, comme l'introduction de bruit. À cause des perturbations, les valeurs des mesures risquent de bouger. Dès lors une étude expérimentale de l'influence du bruit sur les mesures serait intéressante ;
2. Le classement des règles par les mesures d'intérêt semblent aussi une piste de recherche intéressante. En effet, par l'étude de l'ordonnancement des règles, il serait possible d'améliorer les résultats obtenus par l'étude empirique, en montrant à quel degré les mesures d'un même groupe stable ordonnent les règles de la même manière.
3. Une étude future peut consister en l'identification de caractéristiques pertinentes d'un ensemble de données qui permettraient d'indiquer la mesure d'intérêt la mieux appropriée à utiliser ;
4. Une autre direction de recherche qui peut être abordée, est la réalisation d'un cadre applicatif réel, permettant d'accroître le potentiel d'interaction avec l'utilisateur expert du domaine. Il s'agit de rechercher les mesures d'intérêt objectives les plus corrélées avec l'intérêt réel humain, et ce pour un ensemble de données cible ou un domaine bien précis, telles que des données biologiques, médicales [OSK+04], bancaires, etc.

   L'utilisateur peut commencer ses expérimentations au moyen des mesures représentatives que nous lui avons proposées au cours de cette thèse. Un ensemble de règles classées les plus intéressantes par chacune des mesures sélectionnées seront présentées à l'expert du domaine pour les évaluer aussi. Cette confrontation des résultats avec l'avis subjectif de l'utilisateur pourrait nous renseigner davantage sur le comportement des mesures d'intérêt afin de voir quelles mesures sont les mieux adaptées à quels domaines d'application.
5. L'agrégation des mesures et la recherche des règles non dominées. En effet, puisque les mesures d'intérêt appartenant à des groupes différents sont hétérogènes, différentes propositions de règles d'association intéressantes sont suggérées (*une règle peut être jugée pertinente selon une mesure et non pas selon une autre*). Si un utilisateur a par exemple l'intention de tenir compte de toutes les suggestions proposées par ces classes, il peut opter pour la sélection d'une ou plusieurs mesure(s) référente(s) par classe. Afin de sélectionner les règles pertinentes en considérant plusieurs mesures référentes, il serait possible de se baser sur la notion de dominance de Pareto. Une règle r est dite non dominée, si pour toutes les mesures, il n'existe aucune autre règle r' qui la domine, i.e., r' est moins pertinente que r.

ANNEXE A

Liste des mesures d'intérêt

1 : Coefficient de corrélation ou ϕ-coefficient [Pea96]

$$\frac{P(XY)-P(X)P(Y)}{\sqrt{P(X)P(Y)P(\overline{X})P(\overline{Y})}}$$

2 : Cohen ou Kappa [Coh60]

$$2 \times \frac{P(XY)-P(X)P(Y)}{P(X)+P(Y)-2P(X)P(Y)} = 2 \times \frac{P(XY)-P(X)P(Y)}{P(X)P(\overline{Y})+P(\overline{X})P(Y)}$$

3 : Confiance ou précision [AIS93]

$$P(Y/X) = \frac{P(XY)}{P(X)}$$

4 : Confiance causale [Kod01]

$$1 - \frac{1}{2}\left(\frac{1}{P(X)} + \frac{1}{P(\overline{Y})}\right) P(X\overline{Y}) = 1 - \frac{1}{2}\, P(\overline{Y}/X) - \frac{1}{2}\, P(X/\overline{Y}))$$

5 : Confiance centrée ou Pavillon ou Valeur ajoutée [LT04]

$$P(Y/X) - P(Y) = \frac{P(XY)-P(X)P(Y)}{P(X)} = P(\overline{Y}) - P(\overline{Y}/X)$$

6 : Confiance confirmée descriptive ou Ganascia [Gan87]

$$1 - 2 \times P(\overline{Y}/X)$$

7 : Confiance confirmée causale [Kod01]

$$1 - \frac{3}{2}P(\overline{Y}/X) - \frac{1}{2}P(X/\overline{Y}) = 1 - \frac{1}{2}\left(\frac{3}{P(X)} + \frac{1}{P(\overline{Y})}\right) P(X\overline{Y})$$

8 : Confirmation causale [Kod01]

$$P(X) + P(\overline{Y}) - 4 \times P(X\overline{Y})$$

9 : Confirmation descriptive [Kod01]

$$P(X) - 2 \times P(X\overline{Y}) = P(XY) - P(X\overline{Y})$$

10 : Conviction [BMS97]

$$\frac{P(X)P(\overline{Y})}{P(X\overline{Y})}$$

11 : Cosinus ou Ochiai [Och57]

$$\frac{P(XY)}{\sqrt{P(X)P(Y)}}$$

12 : Couverture [Kod01], [OKYY03]

$$P(X)$$

13 : Czekanowski-Dice ou F-mesure [Cze13]

$$\frac{2\times P(XY)}{P(X)+P(Y)} = \frac{2\times P(XY)}{P(XY)+1-P(\overline{X}\ \overline{Y})}$$

14 : Dépendance [Kod01]

$$|P(\overline{Y}) - P(\overline{Y}/X)|$$

15 : Dépendance causale estimée [Kod01]

$$\frac{3}{2} + 2P(X) - \frac{3}{2}P(Y) - \frac{3}{2}P(\overline{Y}/X) - 2P(X/\overline{Y})$$

16 : Dépendance pondérée d'intérêt de Gray et Orlowska [Kod01]

$$\left(\left(\frac{P(XY)}{P(X)P(Y)}\right)^k - 1\right) \times P(XY)^m$$

17 : Facteur bayésien ou Multiplicateur de côte [Jef35]

$$\frac{P(XY)P(\overline{Y})}{P(X\overline{Y})P(Y)}$$

18 : Facteur de certitude ou Loevinger ou Satisfaction [Loe47]

$$\frac{P(Y/X)-P(Y)}{1-P(Y)} = \frac{P(Y/X)-P(Y)}{P(\overline{Y})} = 1 - \frac{P(X\overline{Y})}{P(X)P(\overline{Y})} = \frac{P(X)P(\overline{Y})-P(X\overline{Y})}{P(X)P(\overline{Y})}$$

19 : Fiabilité négative [LFZ99]

$$P(\overline{X}/\overline{Y})$$

20 : Force collective [TKS04]

$$\frac{P(XY)+P(\overline{X}\,\overline{Y})}{P(X)P(Y)+P(\overline{X})P(\overline{Y})} \times \frac{1-P(X)P(Y)-P(\overline{X})P(\overline{Y})}{1-P(XY)-P(\overline{X}\,\overline{Y})}$$

21 : Fukuda [FMMT96]

$$n\left(P(XY) - min_{confiance}\, P(X)\right)$$

22 : Gain informationnel ou Gain d'information [CH90]

$$log_2\left(\frac{P(XY)}{P(X)P(Y)}\right)$$

23 : Gini [TKS04]

$$P(X) \times \left(P(Y/X)^2 + P(\overline{Y}/X)^2\right) + P(\overline{X}) \times \left(P(Y/\overline{X})^2 + P(\overline{Y}/\overline{X})^2\right) - P(Y)^2 - P(\overline{Y})^2$$

24 : Goodman-Kruskal [GK54]

$$\frac{\sum_j max_k P(X_j,Y_k) + \sum_k max_j P(X_j,Y_k) - max_j P(X_j) - max_k P(Y_k)}{2 - max_j P(X_j) - max_k P(Y_k)} \quad \frac{n_{XY}n_{\overline{X}\overline{Y}}/n^2 - n_{\overline{X}Y}n_{X\overline{Y}}/n^2}{n_{XY}n_{\overline{X}\overline{Y}}/n^2 + n_{\overline{X}Y}n_{X\overline{Y}}/n^2}$$

25 : Indice d'implication [LGR81], [LA07]

$$\sqrt{n}\,\frac{P(X\overline{Y})-P(X)P(\overline{Y})}{\sqrt{P(X)P(\overline{Y})}}$$

26 : Indice probabiliste d'écart à l'équilibre (IPEE) [BGBG5a]

$$P\left[\mathcal{N}(0,1) \geq \frac{n_{X\overline{Y}} - n_{XY}}{\sqrt{n_X}}\right]$$

27 : IP3E (IPEE Entropique) [BGBG5b]

$$\sqrt{\frac{1}{2}\Big[((1-h_1(P(X\overline{Y}))^2)\times(1-h_2(P(X\overline{Y}))^2))^{\frac{1}{4}}+1\Big]\times IPEE}$$

with $h_1(t)=-\Big(1-\frac{t}{P(X)}\Big)log_2\Big(1-\frac{t}{P(X)}\Big)-\frac{t}{P(X)}log_2\Big(\frac{t}{P(X)}\Big)\ for\ t\in\Big[0,\frac{P(X)}{2}\Big[,\ else\ h_1(t)=1$,$h_2(t)=-\Big(1-\frac{t}{P(\overline{Y})}\Big)log_2\Big(1-\frac{t}{P(\overline{Y})}\Big)-\frac{t}{P(\overline{Y})}log_2\Big(\frac{t}{P(\overline{Y})}\Big)\ for\ t\in\Big[0,\frac{P(\overline{Y})}{2}\Big[,\ else\ h_2(t)=1$

28 : Indice probabiliste discriminant (IPD) [LA07]

$$P\Big[\mathcal{N}(0,1)\geq II^{CR/B}\Big]$$

où $II^{CR/B}$ indique que *II* est centré-réduit en fonction des valeurs prises par *II* sur l'ensemble des règles extraites.

29 : Information mutuelle [TKS02]

$$\frac{VS(XY)}{-P(X)log_2P(X)-P(\overline{X})log_2P(\overline{X})}$$

30 : Intensité d'implication (II) [Gra79]

$$P\Big[Poisson(nP(X)P(\overline{Y}))\geq P(X\overline{Y})\Big]$$

31 : Intensité d'implication entropique (IIE) [GKCG01]

$$\sqrt{\Big[(1-h_1(P(X\overline{Y}))^2)\times(1-h_2(P(X\overline{Y}))^2)\Big]^{\frac{1}{4}}\times II}$$

32 : Intensité d'implication entropique révisée (IIER) [LVL05]

$$\sqrt{\Big[(1-h_1(P(X\overline{Y}))^2)\times(1-h_2(P(X\overline{Y}))^2)\Big]^{\frac{1}{4}}\times max(2\times II-1;0)}$$

33 : Indice de vraisemblance du lien (IVL) [Ler81]

$$P\Big[Poisson(nP(X)P(Y))<P(XY)\Big]$$

34 : Intérêt ou Lift [BMS97]

$$\frac{P(Y/X)}{P(Y)}=\frac{P(XY)}{P(X)P(Y)}$$

35 : Jaccard [Jac08]

$$\frac{P(XY)}{P(X)+P(Y)-P(XY)}=\frac{P(XY)}{P(X\overline{Y})+P(Y)}$$

36 : J-Mesure [GS89]

$$P(XY)\ log_2\Big(\frac{P(XY)}{P(X)P(Y)}\Big)+P(X\overline{Y})\ log_2\Big(\frac{P(X\overline{Y})}{P(X)P(\overline{Y})}\Big)$$

37 : Klosgen [Klo96]

$$\sqrt{P(XY)}\ \times(P(Y/X)-P(Y))$$

38 : Kulczynski ou Indice d'accord et de désaccord [Kul28]

$$\frac{P(XY)}{P(X\overline{Y})+P(\overline{X}Y)}$$

39 : Laplace [Goo65]

$$\frac{n_{XY}+1}{n_X+2}$$

40 : Leverage [PS91a]

$$P(Y/X) - P(X)P(Y)$$

41 : M_{GK} [Gui00]

$$Si\ P(Y/X) \geq P(Y)\ \ alors\ M_{GK}(X \rightarrow Y) = \frac{P(Y/X)-P(Y)}{1-P(Y)}$$

$$Sinon\ M_{GK}(X \rightarrow Y) = \frac{P(Y/X)-P(Y)}{P(Y)}$$

42 : Moindre contradiction ou Surprise [AK02]

$$\frac{P(XY)-P(X\overline{Y})}{P(Y)}$$

43 : Nouveauté [LFZ99]

$$P(XY) - P(X)P(Y)$$

44 : Pearl [Pea88]

$$P(X)|P(Y/X) - P(Y)|$$

45 : Piatetsky-Shapiro [PS91a]

$$n \times (P(XY) - P(X)P(Y))$$

46 : Précision ou Support causal [TT95], [IIDRS96]

$$P(XY) + P(\bar{X}\bar{Y}) = P(X) + P(\overline{Y}) - 2 \times P(X\overline{Y})$$

47 : Prévalence [OKYY03]

$$P(Y)$$

48 : Q de Yule [Yul27]

$$\frac{P(XY)P(\bar{X}\bar{Y})-P(X\overline{Y})P(\overline{X}Y)}{P(XY)P(\bar{X}\bar{Y})+P(X\overline{Y})P(\overline{X}Y)}$$

49 : Rappel [LFZ99]

$$P(X/Y) = \frac{P(XY)}{P(Y)}$$

50 : Ratio des chances [TKS02]

$$\frac{P(XY)P(\bar{X}\bar{Y})}{P(\overline{X}Y)P(X\overline{Y})}$$

51 : Risque relatif [AMS97]

$$\frac{P(Y/X)}{P(Y/\overline{X})}$$

52 : Sebag-Schoenauer [SS88]

$$\frac{P(XY)}{P(X\overline{Y})}$$

53 : Spécificité [OKYY03]

$$P(\overline{Y}/\overline{X}) = \frac{P(\bar{X}\bar{Y})}{P(\overline{X})}$$

54 : Support ou Indice de Russel et Rao [AIS93]

$$P(XY)$$

55 : Support à sens unique de Yao et Liu (SU) [YZ99]

$$P(Y/X)log_2\frac{P(XY)}{P(X)P(Y)}$$

56 : Support à double sens de Yao et Liu (SD) [YZ99]

$$P(XY)log_2\frac{P(XY)}{P(X)P(Y)}$$

57 : Taux d'exemples et de contre-exemples [VLP06]

$$\frac{P(XY)-P(X\overline{Y})}{P(XY)}$$

58 : Valeur test VT100 [RM08]

$$\phi^{-1}(P[Hypergeometrique(100P(X)P(Y)) \leq P(XY)])$$

59 : Variation du support à double sens de Yao et Liu (VS)

$$P(XY)log_2\frac{P(XY)}{P(X)P(Y)} + P(X\overline{Y})log_2\frac{P(X\overline{Y})}{P(X)P(\overline{Y})} + P(\overline{X}Y)log_2\frac{P(\overline{X}Y)}{P(\overline{X})P(Y)} + P(\bar{X}\bar{Y})log_2\frac{P(\bar{X}\bar{Y})}{P(\overline{X})P(\overline{Y})}$$

60 : Y de Yule [Yul12]

$$\frac{\sqrt{P(XY)P(\bar{X}\bar{Y})}-\sqrt{P(X\overline{Y})P(\overline{X}Y)}}{\sqrt{P(XY)P(\bar{X}\bar{Y})}+\sqrt{P(X\overline{Y})P(\overline{X}Y)}}$$

61 : Zhang [Zha00]

$$\frac{P(XY)-P(X)P(Y)}{max\left\{P(XY)P(\overline{Y}),\ P(Y)P(X\overline{Y})\right\}}$$

ANNEXE B

Typologie des variables

Cette annexe présente la typologie des variables (*ou données*) utilisée dans le chapitre 1 lors de la phase de transformation des variables du processus de l'ECD ainsi que dans les chapitres 2 et 3 concernant les variables d'entrée de notre étude des mesures.

Variable Une variable est une caractéristique ou propriété susceptible d'être possédée ou non par les individus de la population donnée. La taille, le sexe, la couleur des yeux, le nombre de téléviseurs de votre foyer ou encore l'âge sont des variables.

Il existe principalement trois types de variables : les variables logiques, qualitatives et quantitatives. Nous désignons par $X(\Omega)$, l'ensemble des valeurs que peut prendre ces variables.

Variable logique Les variables logiques prennent leur valeurs dans l'ensemble {0,1}, ou {Oui, Non}, ou encore {Vrai, Faux}.

Il est possibles de définir les variables logiques à partir d'une application f_1 de la population Ω dans l'ensemble {0, 1}, tel que :

$$\begin{aligned} f_1 : \Omega &\to \{0,1\} \\ e &\mapsto f_1(e) \end{aligned}$$

où l'individu e représente un élément de Ω. On dit que e est associé à $f_1(e) = 1$ (*respectivement 0*) si l'individu possède la caractéristique de la variable (*respectivement si l'individu ne possède pas la caractéristique de la variable*) chez l'objet e.

Par exemple, les variables "Temps=nuageux" et "Couleur=rouge" sont des variables logiques.

Variable qualitative Les variables qualitatives sont des données auxquelles on peut attribuer une valeur ou une caractéristique. Elles sont représentées par des qualités, telles que le sexe, le programme d'études ou encore l'état civil. Les variables qualitatives s'expriment en modalités [1].

Il existe deux types de variables qualitatives :

1. variables qualitatives nominales : ce sont des variables qui prennent leurs valeurs dans un ensemble fini de modalités dépourvues d'ordre précis. Ce sont seulement des mots dans le désordre. Par exemple, le sexe a 2 modalités possibles (*féminin ou masculin*), la religion (*catholique, protestant, Musulman ...*) ;
2. variables qualitatives ordinales : ce sont des variables qui prennent leurs valeurs dans un ensemble fini de modalités munies d'une relation d'ordre. Par exemple, le degré de satisfaction d'un client pour un produit, qu'on peut noter de -2 (*très mauvais*) à $+2$ (*excellent*), en passant par zéro (*indifférent*).

1. Une modalité est une valeur que peut prendre la variable étudiée.

Variable quantitative Les variables quantitatives sont des données qui peuvent être mesurées (*taille, poids*) ou repérées (*température*). Elles prennent leurs valeurs dans l'ensemble $\Re$ des nombres réels ou dans un intervalle de $\Re$.

Pour les variables quantitatives, il y a encore 2 types de variables différentes :

1. variables discrètes : ce sont des variables qui ne prennent que des valeurs entières ou dénombrables, par exemple le nombre des enfants d'une famille, le nombre de pièces d'un logement ;
2. variables continues : ce sont des variables issues de mesures, dont les valeurs peuvent varier d'aussi peu que l'on voudra dans un intervalle de $\Re$ fini ou infini. Par exemple la taille d'une personne, le poids d'un enfant, les valeurs du pH, etc.

Le *diagramme B.1* résume bien le tout.

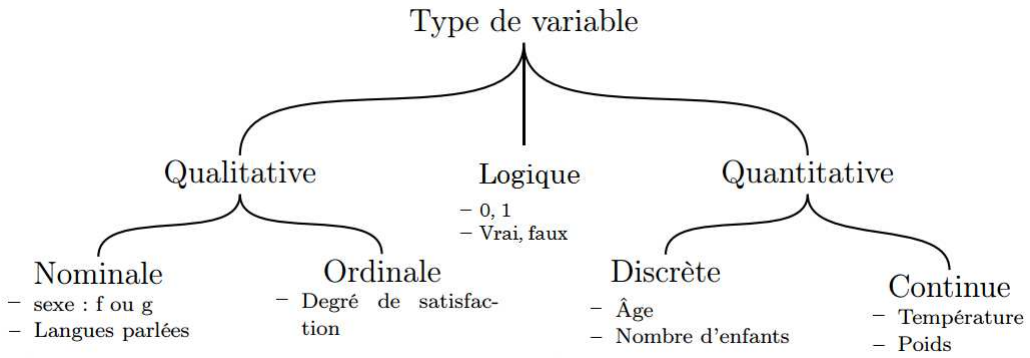

FIGURE *B.1: Diagramme des différents types de variables.*

ANNEXE C

Les critères de la classification non supervisée

Cette annexe rappelle les mesures d'éloignement existantes dans la littérature et définit les principes des deux méthodes de classification non supervisée : *CAH* et *K-moyennes*.

C.1 Mesures d'éloignement

Le clustering repose sur une mesure précise de la similarité/dissemblance des individus que l'on veut regrouper. Une mesure est une formule qui, pour deux individus de l'ensemble de la population, calcule un nombre positif reflétant la proximité/éloignement de ces deux individus. Cette mesure est appelée *distance* ou *métrique*.

La distance entre x et y se note $d(x,y)$. Elle doit satisfaire les propriétés suivantes :

- Symétrique : $d(x,y) = d(y,x)$
- Positive : $d(x,y) \geq 0$
- Nulle d'un individu à lui-même : $d(x,x) = 0$
- Inégalité Triangulaire : $d(x,y) \leq d(x,z) + d(z,y)$

Pour le bon fonctionnement des méthodes de clustering, qui visent à trouver une partition des variables en classes homogènes, il est primordial de choisir une distance pour mesurer la proximité entre les individus. On peut définir de plusieurs manières la distance entre deux individus :

$$distance\ de\ Manhattan : \sum_{i=1}^{n} |x_i - y_i| \tag{C.1}$$

$$distance\ euclidienne : \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{C.2}$$

$$distance\ de\ Minkowski : \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{C.3}$$

$$distance\ de\ Tchebychev : \lim_{p \to \infty} \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} = \sup_{1 \leq i \leq n} |x_i - y_i| \tag{C.4}$$

Parmi ces distances, la distance la plus utilisée entre deux individus est la distance Euclidienne, appelée aussi distance à vol d'oiseau. C'est la racine carrée de la somme des carrés des différences de coordonnées en x et en y.

C.2 Principe de la k-moyenne

L'algorithme *k-moyenne* mis au point par McQueen en 1967 [Mac67], est l'un des algorithmes de clustering les plus connus. Son objectif est de partitionner l'ensemble des individus en k classes, où k est un paramètre fourni par l'utilisateur.

Le principe de l'algorithme des *k-moyennes* est défini dans l'*algorithme 3*.

Algorithme 3: Algorithme des K-moyennes

1 **begin**

2 **Initialisation** : Choisir arbitrairement k centres initiaux des classes $c_1, c_2, \cdots, c_k$ où chaque c_i représente le centre d'une classe C_i

3 **Affectation** : Calcul de la classe en affectant chaque point de la population au centre de gravité dont il est le plus "proche"

4 **Mise à jour des centres** c_i : Calcul des nouveaux centres de chaque classe

5 **Arrêt** : Si aucun changement (atteindre un état de stabilité) dans (3) alors stop, sinon on retourne à (2)

C.3 Principe de la CAH

Les hiérarchies, par la commodité de leur interprétation visuelle, constituent depuis longtemps une forme de classification très populaire [CDG+89]. Généralement, l'utilisateur s'intéresse à l'identification de classes "bien significatives", issues de la hiérarchie. Dès lors, la construction d'une hiérarchie nécessite la connaissance d'une "mesure de ressemblance" entre groupes. Cette mesure est appelée "indice d'agrégation". Les indices (*ou critères*) d'agrégation les plus classiques sont :

- l'indice du **lien maximum**, qui donne la distance maximale entre deux éléments des deux groupes ;
- l'indice du **lien minimum**, qui calcule la distance minimale entre deux éléments des groupes ;
- l'indice de **Ward**, qui calcule l'augmentation de l'inertie intra-classes de la partition lors de la fusion entre les deux groupes.

Ce dernier critère, le critère de Ward [War63], est l'un des plus utilisés, précisément car il permet de minimiser à chaque étape l'inertie intra-classes des partitions obtenues.

Le critère de Ward est défini de la manière suivante :

$$d(C_1, C_2) = \frac{n_1 * n_2}{n_1 + n_2}\, d(G_1, G_2) \quad (\textit{saut de Ward}) \tag{C.5}$$

avec C_1 et C_2 deux classes d'une partition donnée ; n_1 et n_2 sont les effectifs des deux classes et G_1 et G_2 leurs centres de gravité respectifs.

Une fois un indice d'agrégation est choisi, on construit une suite de partitions de moins en moins fines dont les classes forment la hiérarchie cherchée [CDG+89].

Le principe de l'algorithme général de la Classification Ascendante Hiérarchique (*CAH*) est illustré dans l'algorithme 4.

Le dendogramme résultant de l'application de l'algorithme de classification ascendante hiérarchique est une représentation graphique, sous forme d'arbre, des agrégations successives jusqu'à la réunion en une seule classe de tous les individus. La hauteur d'une branche est proportionnelle à l'indice de

Algorithme 4: Algorithme de classification ascendante hiérarchique

```
1 begin
2     Les classes initiales sont les individus
3     Calcul de la matrice de distance M entre chaque couple de clusters
4     repeat
5         Sélection dans M des deux clusters les plus proches C_i et C_j
6         Fusion de C_i et C_j par un cluster C_G plus général
7         Mise à jour de M en calculant la ressemblance entre C_G et les clusters existants
8     until la fusion des 2 derniers clusters, i.e., n'avoir plus qu'une seule classe, qui
      contient tous les individus;
9     Retourner un arbre appelé dendrogramme
```

dissemblance ou distance entre les deux individus regroupés. Dans le cas du saut de Ward, c'est la perte de variance interclasses. Le nombre de classes est obtenu en traçant une ligne horizontale en travers du dendogramme, et en retenant dans la typologie les clusters terminaux qui sont juste au-dessus de cette ligne. En changeant la hauteur de la ligne, on change également le nombre de clusters retenus.

Remarque 1 : Le saut de Ward est la stratégie la plus courante [Tuf05] ; c'est même souvent l'option par défaut dans le cas d'une distance euclidienne entre individus. En effet, ce critère induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse.

Remarque 2 : Les algorithmes précédents sont les techniques de classification "classiques". Ils fonctionnent en prenant en entrée des données représentées soit sous la forme d'un tableau "individus-variables" de largeur fixe, soit sous la forme d'une matrice de similarité ou dissimilarité entre les individus.

Mesures et leurs abréviations			
1	Coefficient de corrélation (Cor)	2	Cohen (Coh)
3	Confiance (Conf)	4	Confiance causale (CConf)
5	Pavillon (Pav)	6	Ganascia (Gan)
7	Confiance confirmée causale (CCC)	8	Confirmation causale (CfmC)
9	Confirmation descriptive (CfmD)	10	Conviction (Conv)
11	Cosinus (Cos)	12	Couverture (Cov)
13	Czekanowski-Dice (CzD)	14	Dépendance (Dep)
15	Dépendance causale (PCD)	16	Dépendance pondérée (DP)
17	Facteur bayésien (FB)	18	Facteur de certitude (FC)
19	Fiabilité négative (FN)	20	Force collective (FCol)
21	Fukuda (Fuk)	22	Gain informationnel (GI)
23	Gini (Gini)	24	Goodman-Kruskal (Good)
25	Indice d'implication (IndImp)	26	Intensité probabiliste d'écart à l'équilibre (IPEE)
27	Intensité probabiliste entropique d'écart à l'équilibre (IP3E)	28	Indice probabiliste discriminant (IPD)
29	Information mutuelle (IM)	30	Intensité d'implication (II)
31	Intensité d'implication entropique (IIE)	32	Intensité d'implication entropique révisée (IIER)
33	Indice de vraisemblance du lien (IVL)	34	Intérêt (Int)
35	Jaccard (Jac)	36	J-mesure (Jmes)
37	Klosgen (Klos)	38	Kulczynski (Kulz)
39	Laplace (Lap)	40	Leverage (Lev)
41	M_{GK} (Mgk)	42	Moindre contradiction (MC)
43	Nouveauté (Nov)	44	Pearl (Pearl)
45	Piatetsky-Shapiro (PS)	46	Précision (Prec)
47	Prévalence (Prev)	48	Q de Yule (YQ)
49	Rappel (Rap)	50	Ratio des chances (RC)
51	Risque relatif (RR)	52	Sebag-Schoenauer (Seb)
53	Spécificité (Spec)	54	Support (Sup)
55	Support à sens unique (SSU)	56	Support à double sens de Yao et Liu (SDS)
57	Taux d'exemples (TEC)	58	VT100
59	Variation support à double sens de Yao et Liu (VS)	60	Y de Yule (YY)
61	Zhang (Zhang)		

TABLE *C.1: Les abréviations des mesures d'intérêt étudiées.*

Bibliographie

[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 94–105. ACM Press, 1998. (cité page 78).

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 207–216, 1993. (cité pages 1, 10, 10, 12, 13, 15, 16, 17, 20, 199 et 202).

[AK02] J. Azé and Y. Kodratoff. Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In *Revue RSTI-ECA (Numéro spécial EGC'2002)*, volume 1, pages 143–154, 2002. (cité pages 21, 29, 37, 69 et 202).

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996. (cité page 16).

[AMS97] K. Ali, S. Manganaris, and R. Srikant. Partial classification using association rules. In *KDD*, pages 115–118, 1997. (cité pages 13 et 202).

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the 20th Intl. Conf. on Very Large Data Bases (VLDB'94)*, pages 478–499. Morgan Kaufmann, 1994. (cité pages 10, 13, 15, 16, 18 et 21).

[AY98] C.C. Aggarwal and P.S. Yu. A new framework for itemset generation. In *PODS*, pages 18–24, 1998. (cité pages 20, 21 et 21).

[Azé03] J. Azé. *Extraction de connaissances à partir de données numériques et textuelles*. Phd thesis, Paris 11, 2003. (cité page 22).

[BA96] R.J. Brachman and T. Anand. The process of knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*, pages 37–57. 1996. (cité page 8).

[Ben73] J.P Benzécri, editor. *L'analyse des données*. Dunod, Paris, 1973. (cité pages 79 et 80).

[Bez81] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New york, USA, 1981. (cité page 126).

[BGBG5a] J. Blanchard, F. Guillet, H. Briand, and R. Gras. Ipee : Indice probabiliste d'écart á l'équilibre pour l'évaluation de la qualité des règles. In *Atelier Qualité des Données et des Connaissances*, pages 26–34, 2005a. (cité pages 28, 30, 30, 40, 43, 48, 49, 55, 62, 62, 121 et 200).

[BGBG5b] J. Blanchard, F. Guillet, H. Briand, and R. Gras. Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. In *Troisièmes rencontres internationales de l'Analyse Statistique Implicative (ASI 05)*, pages 131–137, Palermo, Italy, 2005b. (cité pages 49 et 200).

[BGBG5c] J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *Proceedings of the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, pages 191–200. ENST, 2005c. (cité pages i, 30, 30, 31, 31, 31, 40, 43 et 121).

[BGG+13] R. Belohlavek, D. Grissa, S. Guillaume, E. Mephu Nguifo, and J. Outrata. Boolean factors as a means of clustering interestingness measures of association rules. *Annals of Mathematics and Artificial Intelligence (AMAI)*, 67, 2013. (cité page 145).

[BGGB04] J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. In *Actes des quatrièmes journées Extraction et Gestion des Connaissances*, volume RNTI-E-2 of *Revue des Nouvelles Technologies de l'Information*, pages 287–298, 2004. (cité pages i, 28, 28, 30, 30 et 30).

[BH03] B. Barber and H. J. Hamilton. Extracting share frequent itemsets with infrequent subsets. *Data Min. Knowl. Discov.*, 7(2), 2003. (cité page 21).

[BKGB04] J. Blanchard, P. Kuntz, F. Guillet, and H. Briand. Mesure de qualité des règles d'association par l'intensité d'implication entropique. In *Actes des 4èmes EGC' 04, Clermont-Ferrand*, volume E1 of *Revue des Nouvelles Technologies de l'Information, Numéro spécial Mesures de qualité pour la fouille de données*, pages 33–44. Cépaduès, 2004. (cité page 2).

[BLL12] Y. Le Bras, P. Lenca, and S. Lallich. *Formal framework for the study of algorithmic properties of objective interestingness measures*, volume 24 of *Intelligent Systems Reference Library, Data Mining : Foundations and Intelligent Paradigms*, chapter Data Mining : Foundations and Intelligent Paradigms, pages 77–98. Springer-Verlag, 2012. (cité pages 32, 43, 69, 113 et 121).

[BLLV06] J.-P. Barthélemy, A. Legrain, P. Lenca, and B. Vaillant. Aggregation of valued relations applied to association rule interestingness measures. In *MDAI*, pages 203–214, 2006. (cité page 21).

[BMLL10a] Y. Le Bras, P. Meyer, P. Lenca, and S. Lallich. Mesure de la robustesse de règles d'association. In *QDC 2010 : atelier Qualité des Données et des Connaissances, en conjonction avec Extraction et gestion des connaissances*, pages 27–38, 2010. (cité pages 55 et 69).

[BMLL10b] Y.Le Bras, P. Meyer, P. Lenca, and S. Lallich. A robustness measure of association rules. In *ECML/PKDD (2)*, volume 6322, pages 227–242. Springer, 2010. (cité page 113).

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets : Generalizing association rules to correlations. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 265–276, 1997. (cité pages 2, 2, 21, 21, 21, 50, 52, 199 et 201).

[BPT+00] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *Proc. of the First Intl. Conf. on Computational Logic*, CL'00, pages 972–986, London, UK, 2000. Springer-Verlag. (cité page 21).

[Bra11] Y. Le Bras. *Contribution à l'étude des mesures de l'intérêt des règles d'association et à leurs propriétés algorithmiques*. Phd thèse, Université de Bretagne Sud (UBS), Lab-STICC UMR CNRS 3192., 2011. (cité pages ii, 32, 32, 40, 43, 48, 69, 71, 76, 109, 109, 113, 113, 113, 114, 121, 124, 148, 167, 167, 170, 179, 182, 182, 182 et 183).

[BSYN12] S. Bouker, R. Saidi, S. Ben Yahia, and E. Mephu Nguifo. Ranking and selecting association rules based on dominance relationship. In *IEEE 24th International Conference on Tools with Artificial Intelligence, ICTAI 2012*, pages 658–665, 2012. (cité pages 124 et 144).

[Bun96] W.L. Buntine. Graphical models for discovering knowledge. In *Advances in Knowledge Discovery and Data Mining*, pages 59–82. 1996. (cité page 22).

[BV10] R. Belohlavek and V. Vychodil. Discovery of optimal factors in binary data via a novel method of matrix decomposition. *J. Comput. Syst. Sci.*, 76(1) :3–20, 2010. (cité pages 127, 128, 128, 130, 130, 131, 131, 132, 132, 132 et 133).

[BVW03] T. Brijs, K. Vanhoof, and G. Wets. Defining interestigness for association rules. *International Journal ITA*, 10(4) :370–375, 2003. (cité page 20).

[CA11] E. Chandra and V.P. Anuradha. Article : A survey on clustering algorithms for data in spatial database management systems. *International Journal of Computer Applications*, 24(9) :19–26, June 2011. (cité page 78).

[CDG+89] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données*. Dunod, 1989. (cité pages 80, 208 et 208).

[Cel89] G. Celeux. *Classification automatique des données : environnement statistique et informatique*. Dunod, 1989. (cité page 78).

[CFE05] D. R. Carvalho, A. A. Freitas, and N. Ebecken. Evaluating the correlation between objective rule interestingness measures and real human interest. In *Proc. of Principles of Data Mining and Knowledge Discovery*, PKDD '05, pages 453–461. Springer, 2005. (cité pages ii, 21, 38, 38, 41, 41, 44, 148 et 192).

[CH90] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1) :22–29, 1990. (cité page 200).

[CHY96] M.S. Chen, J. Han, and P.S Yu. Data mining : An overview from a database perspective. *IEEE Trans. on Knowl. and Data Eng.*, 8(6) :866–883, 1996. (cité page 78).

[CMV04] G. Cleuziou, L. Martin, and C. Vrain. Poboc : An overlapping clustering algorithm, application to rule-based classification and textual data. In *Proc. of the 16th European Conf. on Artificial Intelligence (ECAI-04)*, pages 440–444, 2004. (cité page 126).

[CN06] O. Couturier and E. Mephu Nguifo. Une approche anthropocentrée interactive pour l'aide à la décision en marketing bancaire. In *Proc. of the 18th Intl. Conf. of the Association Francophone d'Interaction Homme-Machine, IHM'2006*, volume 133 of *ACM International Conference Proceeding Series*, pages 253–256, 2006. (cité page 154).

[Coh60] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1) :37–46, 1960. (cité pages 49 et 199).

[CR93] C. Carpineto and G. Romano. GALOIS : An order-theoretic approach to conceptual clustering. In *Proceedings of the 10th International Conference on Machine Learning (ICML'90)*, pages 33–40, 1993. (cité page 78).

[CR06] A. Ceglar and J.F. Roddick. Association mining. *ACM Comput. Surv.*, 38(2), 2006. (cité page 10).

[CS96] P. Cheeseman and J. Stutz. Advances in knowledge discovery and data mining. chapter Bayesian classification (AutoClass) : theory and results, pages 153–180. American Association for Artificial Intelligence, 1996. (cité page 78).

[CYS03] R. Chan, Q. Yang, and Yi-D. Shen. Mining high utility itemsets. In *ICDM*, pages 19–26, 2003. (cité page 23).

[Cze13] J. Czekanowski. Zarys metod statystycznych (die grundzuge der statischen metoden. 1913. (cité pages 49 et 199).

[Czy96] A. Czyzewski. Mining knowledge in noisy audio data. In *Proc. of the 2nd ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '96, pages 220–225. AAAI Press, 1996. (cité page 13).

[Did71] E. Diday. Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de Statistique Appliquée*, XIX(2) :19–33, 1971. (cité page 79).

[Did82] E. Diday. *Éléments d'analyse de données*. Dunod décision. Dunod, 1982. (cité page 88).

[Did86] E. Diday. *Orders and Overlapping Clusters by Pyramids, Multidimentional Data Analysis*. dswo Press, Leiden, 1986. (cité page 126).

[DRT07] J. Diatta, H. Ralambondrainy, and A. Totohasina. Towards a unifying probabilistic implicative normalized quality measure for association rules. In *Quality Measures in Data Mining*, pages 237–250. 2007. (cité page 57).

[Dun73] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3) :32–57, 1973. (cité page 126).

[EKX95] M. Ester, H.-P. Kriegel, and X. Xu. A database interface for clustering in large spatial databases. In *Proc. ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '95, pages 94–99. AAAI Press, 1995. (cité page 78).

[Fen07] D.J. Feno. *Mesures de qualité des règles d'association : normalisation et caractérisation des bases.* Phd thesis, Université de La Réunion., 2007. (cité pages ii, 31, 31, 31, 43, 48, 48, 76, 121, 124 et 148).

[FMMT96] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules : Scheme, algorithms, and visualization. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 13–23, 1996. (cité pages 49 et 200).

[FPSM91] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases : An overview. In *Knowledge Discovery in Databases*, pages 1–30. AAAI/MIT Press, 1991. (cité pages 1, 8 et 8).

[FPsS96] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17 :37–54, 1996. (cité page 21).

[FPSSU96] M. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. (cité pages 8, 8, 9 et 21).

[Fre99] A. A. Freitas. On rule interestingness measures. *Knowledge-Based Systems*, 12 :309–315, 1999. (cité pages 2, 2, 2, 20, 21, 22, 22, 27, 56 et 57).

[GAB+96] R. Gras, S. Ag. Almouloud, M. Bailleuil, A. Larher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina. L'implication statistique, nouvelle méthode exploratoire de données. 1996. (cité page 53).

[GAIM00] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani. Mining the stock market (extended abstract) : which measure is best ? In *Proc. of the sixth ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, KDD '00, pages 487–496, New York, NY, USA, 2000. ACM. (cité page 22).

[Gan87] J. G. Ganascia. Charade : A rule system learning system. In *Proc. of the tenth International Jointed Conference in Artificial Intelligence (IJCAI*, pages 345–347, 1987. (cité pages 49 et 199).

[GCB+04] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, and P. Peter. *Mesures de qualité pour la fouille de données*, chapter Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, pages 3–32. Cépaduès éditions, 2004. (cité pages 29 et 67).

[GGM] S. Guillaume, D. Grissa, and E. Mephu Nguifo. Propriétés des mesures d'intérêt pour l'extraction des règles. Technical Report RR-09-10, Rapport de recherche LIMOS, 31 décembre 2009. (cité page 49).

[GGM10] S. Guillaume, D. Grissa, and E. Mephu Nguifo. Propriétés des mesures d'intérêt pour l'extraction des règles. In *Actes de l'atelier QDC de la conférence EGC*, pages 15–28, Hammamet, Tunisie, 2010. (cité pages 49 et 116).

[GH07] L. Geng and H. J. Hamilton. Choosing the right lens : Finding what is interesting in data mining. In Fabrice Guillet and Howard J. Hamilton, editors, *Quality Measures in Data Mining.*, volume 43 of *Studies in Computational Intelligence*, pages 3–24. Springer, 2007. (cité pages i, 2, 21, 28, 28, 30, 40, 43, 48, 48, 55, 58, 76, 121 et 148).

[GK54] L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *J. Am. Stat. Assoc.*, 49 :732–764, 1954. (cité page 200).

[GKCG01] R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In *EGC*, volume 1 of *Extraction des Connaissances et Apprentissage*, pages 69–80, 2001. (cité pages 21, 27, 48, 49, 64, 64, 68 et 201).

[Goo65] I.J. Good. The estimation of probabilities : An essay on modern bayesian methods, 1965. (cité page 202).

[Gra79] R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. Phd thesis, Université de Rennes I, France, 1979. (cité pages 53, 67 et 201).

[GS89] R.M. Goodman and P. Smyth. The induction of probabilistic rule sets the itrule algorithm. In *Proc. of the sixth international workshop on Machine learning*, pages 129–132. Morgan Kaufmann Publishers Inc., 1989. (cité page 201).

[GSM94] G.Piatetsky-Shapiro and C.J. Matheus. The interestingness of deviations. In *Knowledge Discovery and Data Mining Workshop*, KDD '94, pages 25–36. AAAI Press, 1994. (cité page 13).

[Gué06] S. Guérif. *Réduction de dimension en apprentissage numérique non supervisé*. 2006. (cité page 88).

[Gui00] S. Guillaume. *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales.* Phd thesis, Université de Nantes, France, 2000. (cité pages 25, 31, 49, 53, 53 et 202).

[GW97] B. Ganter and R. Wille. *Formal Concept Analysis : Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 1997. (cité page 131).

[GZ01] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In *ICDM*, pages 163–170, 2001. (cité page 152).

[Har75] J.A. Hartigan. *Clustering Algorithms*. John Wiley & Sons Inc., New York, 1975. (cité page 77).

[HB99] S. Hettich and S.D. Bay. *The UCI KDD Archive [http ://kdd.ics.uci.edu]*. Irvine, CA : University of California, Department of Information and Computer Science, 1999. (cité page 152).

[HC06] C. Hébert and B. Crémilleux. Optimized rule mining through a unified framework for interestingness measures. In *Proc. of the 8th Intl. Conf. on Data Warehousing and Knowledge Discovery*, DaWaK '06, pages 238–247, Berlin, Heidelberg, 2006. Springer-Verlag. (cité pages ii, 37, 37, 41, 44 et 192).

[HC07] C. Hébert and B. Crémilleux. A Unified View of Objective Interestingness Measures. In *5th International Conference on Machine Learning and Data Mining (MLDM'07)*, pages 533–547. Springer-Verlag, 2007. (cité pages ii, 37, 38, 41, 41, 44, 116 et 192).

[HGB05a] X.-H. Huynh, F. Guillet, and H. Briand. Clustering interestingness measures with positive correlation. In *Proceedings ICEIS (2)*, pages 248–253, 2005. (cité pages ii, 34, 34, 40, 41, 41, 44, 116, 148, 168 et 192).

[HGB05b] X.-H. Huynh, F. Guillet, and H. Briand. A data analysis approach for evaluating the behavior of interestingness measures. In *Proc. of the 8th Intl. Conf. on Discovery Science*, DS '05, pages 330–337. Springer-Verlag, 2005. (cité pages 44 et 192).

[HGB06] X.-H. Huynh, F. Guillet, and H. Briand. Arqat : plateforme exploratoire pour la qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (EGC : Etat et perspectives)*, RNTI-E-5., 2006. (cité page 34).

[HGB+07] X.-H. Huynh, F. Guillet, J. Blanchard, P. Kuntz, H. Briand, and R. Gras. A graph-based clustering approach to evaluate interestingness measures : A tool and a comparative study. In *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*, pages 25–50. Springer, 2007. (cité pages 2, 21, 34, 44, 48, 48, 76, 109, 109, 110, 110, 148, 164, 179, 180, 181, 181 et 192).

[HGN00] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining a general survey and comparison. volume 2, pages 58–64, New York, NY, USA, 2000. ACM. (cité page 1).

[HH00] R.J. Hilderman and H.J. Hamilton. Applying objective interestingness measures in data mining systems. In *Proc. of the 4th European Symposium on Principles of Data Mining and Knowledge Discovery*, PKDD '00, pages 432–439, 2000. (cité page 22).

[HH01] R. J. Hilderman and H. J. Hamilton. *Knowledge Discovery and Measures of Interest*, volume 638 of *The Springer International Series in Engineering and Computer Science*. Kluwer Academic Publishers, Norwell, MA, USA, 2001. (cité pages 2, 21 et 148).

[HKM+96] K. Hätönen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen. Knowledge discovery from telecommunication network alarm databases. In *ICDE*, pages 115–122, 1996. (cité page 13).

[HLSL00] F. Hussain, H. Liu, E. Suzuki, and H. Lu. Exception rule mining with a relative interestingness measure. In *Proc. of Pacific Asia Conference on Knowledge Discovery in DataBases*, PAKDD '00, pages 86–97, 2000. (cité page 21).

[HMS01] D. Hand, H. Mannila, and P. Smyth. Principles of data mining. MIT Press, Cambridge, MA, 2001. (cité page 8).

[Hue09] R.A. Huebner. Diversity-based interestingness measures for association rule mining. In *Proc. of ASBBS*, volume 16, page 1, Las Vegas, 2009. (cité page 13).

[Huy06] X.-H. Huynh. *Interestingness Measures for Association Rules in a KDD Process : Postprocessing of Rules with ARQAT Tool*. Phd thesis, Université de Nantes, France, 2006. (cité pages ii, 31, 31, 31, 40, 43, 121 et 124).

[HW01] H. Hofmann and A. Wilhelm. Visual comparison of association rules. *Computational Statistics*, 16(3) :399–415, 2001. (cité page 1).

[HY02] C. Hacène and T. Yannick. Fouille de textes par combinaison de règles d'association et d'indices statistiques. In *1er Colloque International sur la Fouille de Textes - CIFT'2002*, Hammamet, Tunisie, 2002. (cité page 13).

[HYN+12] L. Haibing, H. Yuan, S.W. Nick, W. Fei, and T. Hanghang. Overlapping clustering with sparseness constraints. In *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops*, pages 486–494. IEEE Computer Society, 2012. (cité page 125).

[HZ10] M.J. Heravi and R. Zaïane. A study on interestingness measures for associative classifiers. In *Proc. of the 2010 ACM SAC, Sierre, Switzerland*, pages 1039–1046. ACM, 2010. (cité pages ii, ii, 32, 32, 36, 36, 40, 43, 44, 48, 76, 109, 109, 111, 111, 111, 121, 124, 148 et 192).

[JA99] R. J. Bayardo Jr. and R. Agrawal. Mining the most interesting rules. In *KDD*, pages 145–154, 1999. (cité pages 20 et 21).

[Jac08] P. Jaccard. Nouvelles recherches sur la distribution florale. In *Bulletin de la Société Vaudense des Sciences Naturelles.*, number 44, pages 223–270, 1908. (cité page 201).

[JBC13] A.d. JimÃľnez, F. Berzal, and J.C Cubero. Interestingness measures for association rules within groups. *Journal Intelligent Data Analysis*, 17(2) :195–215, 2013. (cité pages 2 et 78).

[JD88] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall, New Jersey, USA, 1988. (cité page 77).

[Jef35] H. Jeffreys. Some tests of significance treated by the theory of probability. In *Proc. of the Cambridge Philosophical Society*, pages 203–222, 1935. (cité pages 52 et 200).

[JK13] A. Joshi and R. Kaur. A review : Comparative study of various clustering techniques in data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(3), 2013. (cité page 78).

[JMF99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering : a review. *ACM Comput. Surv.*, 31(3) :264–323, 1999. (cité page 78).

[JS01] S. Jaroszewicz and D.A. Simovici. A general measure of rule interestingness. In *Proc. of Principles of Data Mining and Knowledge Discovery*, PKDD '01, pages 253–265, 2001. (cité page 21).

[KK04] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph. In *SDM*, 2004. (cité page 13).

[KK06] S. Kotsiantis and D. Kanellopoulos. Association rules mining : A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1) :71–82, 2006. (cité page 11).

[Klo96] W. Klosgen. Advances in knowledge discovery and data mining. chapter Explora : a multipattern and multistrategy discovery assistant, pages 249–271. 1996. (cité page 201).

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. of the third Intl. Conf. on Information and knowledge management*, pages 401–407, New York, NY, USA, 1994. ACM. (cité pages 10, 20 et 21).

[KMT97] M. Klemettinen, H. Mannila, and H. Toivonen. A data mining methodology and its application to semi-automatic knowledge acquisition. In *DEXA Workshop*, pages 670–677, 1997. (cité page 13).

[Kod99] Y. Kodratoff. Quelques contraintes symboliques sur le numérique en ecd et ect. In *SFDS*, pages 183–188, Grenoble,France, 1999. (cité pages 13 et 57).

[Kod01] Y. Kodratoff. Comparing machine learning and knowledge discovery in databases : An application to knowledge discovery in texts. In *Machine Learning and Its Applications*, volume 2049, pages 1–21. Springer-Verlag New York, 2001. (cité pages 49, 199, 199, 199, 199, 199, 200, 200 et 200).

[Kon95] I. Kononenko. On biases in estimating multi-valued attributes. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI '95, pages 1034–1040. Morgan Kaufmann Publishers Inc., 1995. (cité page 22).

[KR90] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis*. John Wiley and Sons, 1990. (cité pages 78 et 79).

[KS96] M. Kamber and R. Shinghal. Proposed interestingness measure for characteristic rules. In *AAAI/IAAI, Vol. 2*, pages 1393–1393, 1996. (cité pages 8 et 21).

[KSH01] T. Kohonen, M.R. Schroeder, and T.S. Huang, editors. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001. (cité page 79).

[Kul28] S. Kulczynski. Die p anzenassoziationen der pieninen. pages 57–203, 1928. (cité pages 49 et 201).

[LA07] I.-C. Lerman and J. Azé. *A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link*, volume 43 of Studies in Computational Intelligence, pages 207–236. Springer, 2007. (cité pages 52, 200 et 201).

[Lal02] S. Lallich. *Mesure et validation en extraction des connaissances à partir des données*. Habilitation à diriger des recherches, Université Lyon 2, 2002. (cité pages 21, 29, 58 et 67).

[Lel93] A. Lelu. *Modèles neuronaux pour l'analyse de données documentaires et textuelles*. Phd thèse, Université de Paris VI, 1993. (cité page 125).

[Ler70] I.-C. Lerman. Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité). *Mathématiques et sciences humaines*, 32 :5–15, 1970. (cité pages 49, 91 et 171).

[Ler81] I.-C. Lerman. *Classification et analyse ordinale des données*. Dunod, 1981. (cité page 201).

[LFZ99] N. Lavrac, P. Flach, and B. Zupan. Rule evaluation measures : A unifying view. In *Proc. of the 9th Intl. Workshop on Inductive Logic Programming*, ILP '99, pages 174–185. Springer-Verlag, 1999. (cité pages 21, 49, 52, 200, 202 et 202).

[LGR81] I.C. Lerman, R. Gras, and H. Rostam. Élaboration et évaluation d'un indice d'implication pour des données binaires. i. *Mathématiques et Sciences Humaines*, 74 :5–35, 1981. (cité pages 52 et 200).

[LH96] B. Liu and W. Hsu. Post-analysis of learned rules. In *AAAI/IAAI, Vol. 1*, pages 828–834, 1996. (cité page 23 et 23).

[LHCM00] B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5) :47–55, 2000. (cité page 22).

[IIDRS96] B. De la Iglesia, J.C.W. Debuse, and V.J. Rayward-Smith. Discovering knowledge in commercial databases using modern heuristic techniques. In *Proc. of the ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '96, pages 44–49, 1996. (cité page 202).

[LL99] H. Liu and H. Lu. Mining weak rules. In *COMPSAC*, pages 309–310, 1999. (cité pages 21 et 23).

[LMP$^+$03] P. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich. Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information, Mesures de Qualité pour la Fouille de Données*, 1 :123–134, 2003. (cité pages 2, 2, 27, 27, 55, 61 et 61).

[LMPV03] P. Lenca, P. Meyer, P. Picouet, and B. Vaillant. Aide multicritère à la décision pour évaluer les indices de qualité des connaissances. In *EGC*, pages 271–282, 2003. (cité pages 1, 2, 28, 35, 48, 55 et 55).

[LMVL04] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. A multicriteria decision aid for interestingness measure selection. Technical report, LUSSI, Université du Luxembourg, ERIC - Equipe de recherche en ingénierie des connaissances (Université de Lyon 2), may 2004. Technical Report LUSSI-TR-2004-01-EN. (cité pages 22, 29 et 35).

[LMVL08] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. On selecting interestingness measures for association rules : User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2) :610–626, 2008. (cité pages 21, 31, 164 et 183).

[Loe47] J. Loevinger. A systematic approach to the construction and evaluation of tests of ability. In *Psychological monographs*, 1947. (cité pages 53 et 200).

[LQ12] M. Liu and J. Qu. Mining high utility itemsets without candidate generation. In *Proc. of the 21st ACM Intl. Conf. on Information and knowledge management*, CIKM '12, pages 55–64. ACM, 2012. (cité page 23).

[LR10] M.-J. Lesot and M. Rifqi. Order-based equivalence degrees for similarity and distance measures. In *IPMU'10*, pages 19–28. Springer-Verlag, 2010. (cité pages 76, 109, 109, 114, 115, 115, 115, 116 et 167).

[LT04] S. Lallich and O. Teytaud. Évaluation et validation de mesures d'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information.*, RNTI-E-1. Cépaduès.(2) :193–217, 2004. (cité pages i, 2, 21, 22, 27, 27, 30, 30, 40, 43, 48, 48, 48, 50, 50, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 56, 56, 57, 60, 61, 61, 64, 67, 67, 68, 68, 76, 121, 148 et 199).

[LVL05] S. Lallich, B. Vaillant, and P. Lenca. Parametrised measures for the evaluation of association rule interestingness. In *The XIth Intl. Symp. on Applied Stochastic Models and Data Analysis*, pages 220–229, 2005. (cité page 201).

[LVML07] P. Lenca, B. Vaillant, P. Meyer, and S. Lallich. Association rule interestingness measures : Experimental and theoretical studies. In *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*, pages 51–76. Springer, 2007. (cité pages 35, 40, 41, 44, 166, 170 et 192).

[Mac67] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967. (cité pages 79, 80 et 208).

[McG05] K. McGarry. A survey of interestingness measures for knowledge discovery. *Knowledge Eng. Review*, 20(1) :39–61, 2005. (cité pages 2, 8 et 21).

[MFM+98] Y. Morimoto, T. Fukuda, H. Matsuzawa, T. Tokuyama, and K. Yoda. Algorithms for mining association rules for binary segmentations of huge categorical databases. In *Prod. of the 24rd International Conference on Very Large Data Bases*, VLDB '98, pages 380–391, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. (cité page 38).

[MG07] M. Maddouri and J. Gammoudi. On semantic properties of interestingness measures for extracting rules from data. In *Proc of the 8th international conference on Adaptive and Natural Computing Algorithms, Part I*, ICANNGA '07, pages 148–158. Springer-Verlag, 2007. (cité pages i, 29, 29, 30, 43 et 121).

[MM04] K. McGarry and J. Malone. Analysis of rules discovered by the data mining process. In *Applications and Science in Soft Computing Series : Advances in Soft Computing*, pages 219–224. Springer, 2004. (cité page 22).

[MTV94] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In *Knowledge Discovery and Data Mining Workshop*, KDD '94, pages 181–192, 1994. (cité page 16).

[MYGS91] M. McLeach, P. Yao, M. Garg, and T. Stirtzinger. Discovery of medical diagnostic information : An overview of methods and results. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 477–490. AAAI Press, 1991. (cité page 13).

[Och57] A. Ochiai. Zoogeographic studies on the soleoid shes found in japan and its neighbouring regions, bull. 22 :526–530, 1957. (cité pages 49 et 199).

[OKYY03] M. Ohsaki, S. Kitaguchi, H. Yokoi, and T. Yamaguchi. Investigation of rule interestingness in medical data mining. In *Active Mining*, pages 174–189, 2003. (cité pages 199, 202 et 202).

[OO98] C. Ordonez and E. Omiecinski. Image mining : A new approach for data mining. Technical report, Georgia Institute of Technology, Atlanta, USA, 1998. (cité page 13 et 13).

[OSK+04] M. Ohsaki, Y. Sato, S. Kitaguchi, H. Yokoi, and T. Yamaguchi. Comparison between objective interestingness measures and real human interest in medical data mining. In *Proc. of the 17th Intl. Conf. on Innovations in Applied Artificial Intelligence*, IEA/AIE'2004, pages 1072–1081. Springer Verlag Inc, 2004. (cité pages ii, 38, 38, 41, 41, 44, 192 et 197).

[Pat10] B. Patrice. Classifications en classes recouvrantes ou non, et leurs dissimilarités. *Mathématiques et Sciences Humaines*, 2(190) :59–87, 2010. (cité page 126).

[PBTL99] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proc. of the 7th ICDT'99*, pages 398–416, London, UK, 1999. Springer-Verlag. (cité page 20).

[Pea96] K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. *Philosophical Transactions of The Royal Society A : Mathematical, Physical and Engineering Sciences*, 187 :253–318, 1896. (cité page 199).

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901. (cité page 79).

[Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. (cité page 202).

[PKSL06] M. Plasse, N. Niang Keita, G. Saporta, and L. Leblond. Une comparaison de certains indices de pertinence des règles d'association. In *EGC*, volume RNTI-E-6 of *Revue des Nouvelles Technologies de l'Information*, pages 561–568. Cépaduès-Éditions, January 2006. (cité pages ii, 35, 35, 40, 44 et 192).

[PMS97] M. J. Pazzani, S. Mani, and W.R. Shankle. Comprehensible knowledge-discovery in databases. In *Proc.of the 19th Intl. Conf. of the Cognitive Science society (COGSCI'97)*, pages 596–601. Lawrence Erlbaum, August 1997. (cité page 13).

[PS91a] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248, Cambridge, Mass., 1991. AAAI/MIT Press. (cité pages i, 2, 20, 20, 21, 21, 26, 26, 26, 30, 38, 41, 43, 48, 48, 55, 55, 55, 55, 55, 57, 60, 60, 61, 62, 63, 121, 202 et 202).

[PS91b] G. Piatetsky-Shapiro. Knowledge discovery in real databases : A report on the ijcai-89 workshop. *AI Magazine*, 11(5) :68–70, 1991. (cité pages 2 et 8).

[PT98a] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *ACM Intl. Conf. on Knowledge Discovery and Data Mining*, pages 94–100. AAAI Press, 1998. (cité page 20).

[PT98b] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '98, pages 94–100. AAAI Press, 1998. (cité page 22).

[RFdSG04] W. Romão, A.A. Freitas, and I.M. de Souza Gimenes. Discovering interesting knowledge from a science and technology database with a genetic algorithm. *Appl. Soft Comput.*, 4(2) :121–137, 2004. (cité page 22).

[RM08] R. Rakotomalala and A. Morineau. *The TVpercent principle for the counterexamples statistic*, volume 127 of *Statistical Implicative Analysis : theory and applications*, pages 449–462. Springer, Heidelberg, Germany, 2008. (cité pages 49 et 203).

[Sah99] S. Sahar. Interestingness via what is not interesting. In *Proc. of the 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 332–336, 1999. (cité pages 23 et 185).

[SBK03] E. Segal, A. Battle, and D. Koller. Decomposing gene expression into cellular processes. In *Proc. of 8th Pacific Symposium on Biocomputing (PSB)*, pages 89–100, 2003. (cité page 125).

[SG91] P. Smyth and R.M. Goodman. Rule induction using information theory. In *Knowledge Discovery in Databases, KDD*, pages 159–176. AAAI/MIT Press, 1991. (cité pages 2, 22 et 37).

[SKR10] A. Surana, U. Kiran, and P. K. Reddy. Selecting a right interestingness measure for rare association rules. In *Proc. of the 16th Intl. Conf. on Management of Data*, pages 115–124, Nagpur, India, 2010. Computer Society of India. (cité pages ii, 39, 39, 41, 44, 148 et 192).

[SM02] J. Sese and S. Morishita. Answering the most correlated n association rules efficiently. In *Proc. of Principles of Data Mining and Knowledge Discovery*, PKDD '02, pages 410–422. Springer-Verlag., 2002. (cité pages 1 et 48).

[SS88] M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *J. Boose, B. Gaines, et M. Linster, editors, Proc. of the European Knowledge Acquisition Workshop (EKAW'88)*, 1988. (cité pages 54 et 202).

[ST95] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proc. of the ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '95, pages 275–281, 1995. (cité pages 22 et 23).

[ST96] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE transactions on knowledge and data engineering*, 8 :970–974, 1996. (cité pages 2, 2, 8, 21 et 23).

[Suz08] E. Suzuki. Pitfalls for categorizations of objective interestingness measures for rule discovery. In *Statistical Implicative Analysis*, volume 127 of *Studies in Computational Intelligence*, pages 383–395. Springer, 2008. (cité pages 37, 116, 128 et 148).

[Suz09a] E. Suzuki. Compression-based measures for mining interesting rules. In *Proc. of the 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems : Next-Generation Applied Intelligence*, IEA/AIE '09, pages 741–746. Springer-Verlag, 2009. (cité pages ii, 37, 37 et 37).

[Suz09b] E. Suzuki. Negative encoding length as a subjective interestingness measure for groups of rules. In *Proc. of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '09, pages 220–231. Springer-Verlag, 2009. (cité pages ii, 37, 37, 37, 37, 44 et 192).

[SW85] A. Shoshani and H.K.T. Wong. Statistical and scientific database issues. *IEEE Trans. Software Eng.*, 11(10) :1040–1047, 1985. (cité page 13).

[TGCTB13] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, 27, 2013. (cité pages 1, 2 et 184).

[TKS02] P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proc. of the 8th ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '02, pages 32–41, 2002. (cité pages i, ii, 2, 22, 26, 26, 30, 31, 36, 36, 40, 43, 44, 48, 48, 48, 56, 65, 65, 66, 76, 121, 148, 192, 201 et 202).

[TKS04] P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4) :293–313, 2004. (cité pages 2, 21, 21, 55, 55, 55, 55, 55, 55, 55, 56, 58, 200 et 200).

[TS06] R. Tamir and Y. Singer. On a confidence gain measure for association rule discovery and scoring. *Very Large Data Bases, VLDB*, 15(1) :40–52, 2006. (cité page 21).

[TT95] S. Tsumoto and H. Tanaka. Automated discovery of functional components of proteins from amino-acid sequences based on rough sets and change of representation. In *Proc. of the ACM Intl. Conf. on Knowledge Discovery and Data Mining*, KDD '95, pages 318–324. AAAI Press, 1995. (cité page 202).

[Tuf05] S. Tufféry. *Data Mining et statistique décisionnelle : L'intelligence dans les bases de données*. Ed. Technip, 2005. (cité pages 80 et 209).

[Vai02] B. Vaillant. Évaluation de connaissances : le problème du choix d'une mesure de qualité en extraction de connaissances à partir des données. Master's thesis., École Nationale Supérieure des Télécommunications de Bretagne., 2002. (cité page 35).

[Vai06] B. Vaillant. *Mesurer la qualité des règles d'association : études formelles et expérimentales*. Phd thesis, Université de Bretagne sud, France, 2006. (cité pages i, 30, 30, 40, 40, 43, 48, 48, 55, 56, 64, 67, 71, 76, 76, 84, 100, 109, 109, 109, 109, 116, 121, 124, 148, 148, 179, 179 et 179).

[VLL04] B. Vaillant, P. Lenca, and S. Lallich. Étude expérimentale de mesures de qualités de règles d'association. In *Actes des 4èmes EGC' 04, Clermont-Ferrand*, pages 341–352. Cépaduès, 2004. (cité pages ii, 35, 35, 41, 41, 44, 148, 164 et 192).

[VLP06] B. Vaillant, S. Lallich, and P.Lenca. Modeling of the counter-examples and association rules interestingness measures behavior. In *DMIN*, pages 132–137, 2006. (cité page 203).

[War63] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963. (cité pages 80, 81 et 208).

[WF00] I.H. Witten and E. Frank. *Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco, 2000. (cité page 153).

[WKQ+07] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, J.G. McLachlan, A. Ng, B. Liu, S.P. Yu, Z-H. Zhou, M. Steinbach, J.D. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1) :1–37, 2007. (cité pages 16 et 149).

[XHD+05] H. Xiong, X. He, C.H.Q. Ding, Y. Zhang, V. Kumar, and S.R. Holbrook. Identification of functional modules in protein complexes via hyperclique pattern discovery. In *Pacific Symposium on Biocomputing*, 2005. (cité page 13).

[XL07] Y. Xu and Y. Li. Generating concise association rules. In *CIKM*, pages 781–790, 2007. (cité page 20).

[YH06] H. Yao and H.J. Hamilton. Mining itemset utilities from transaction databases. *Data Knowl. Eng.*, 59(3) :603–626, 2006. (cité page 23).

[Yul12] G.U. Yule. On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society*, 75(6) :579–652, 1912. (cité page 203).

[Yul27] G.U. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London Series. Series A, Containing papers of a Mathematical or Physical Character*, 226 :267–298, 1927. (cité page 202).

[YZ99] Y. Yao and N. Zhong. An analysis of quantitative measures associated with rules. In *Proc. of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, PAKDD '99, pages 479–488. Springer-Verlag, 1999. (cité pages 54, 203 et 203).

[ZAB11] D.A. Zighed, R. Abdesselam, and A. Bounekkar. Équivalence topologique entre mesures de proximité. In *EGC*, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 53–64. Hermann-Éditions, 2011. (cité pages 76, 109, 109, 114, 115, 115 et 116).

[Zak00] M. J. Zaki. Generating non-redundant association rules. In *ACM Intl. Conf. on Knowledge Discovery and Data Mining*, pages 34–43. ACM Press, 2000. (cité page 1).

[Zha00] T. Zhang. Association rules. In *Proc. 4th Pacific-Asia Conference Knowledge Discovery and Data Mining*, pages 7, 47, 2000. (cité pages 21 et 203).

[ZHL+98] O.R. Zaïane, J. Han, Ze-N. Li, S. Han Seng Chee, and J. Chiang. Multimediaminer : A system prototype for multimedia data mining. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 581–583, 1998. (cité page 13).

[ZK01] Y. Zhao and G. Karypis. Criterion Functions for Document Clustering : Experiments and Analysis. Technical report, University of Minnesota, Minneapolis, 2001. (cité page 22).

[ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny. Birch : An efficient data clustering method for very large databases. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, SIGMOD '96, pages 103–114, 1996. (cité page 78).

[ZTS04] M. Zaman Ashrafi, D. Taniar, and K. A. Smith. A new approach of eliminating redundant association rules. In *DEXA*, pages 465–474. Springer, 2004. (cité page 1).

Résumé :

La recherche de règles d'association intéressantes est un domaine important et actif en fouille de données. Puisque les algorithmes utilisés en extraction de connaissances à partir de données (*ECD*), ont tendance à générer un nombre important de règles, il est difficile à l'utilisateur de sélectionner par lui-même les connaissances réellement intéressantes. Pour répondre à ce problème, un post-filtrage automatique des règles s'avère essentiel pour réduire fortement leur nombre. D'où la proposition de nombreuses mesures d'intérêt dans la littérature, parmi lesquelles l'utilisateur est supposé choisir celle qui est la plus appropriée à ses objectifs. Comme l'intérêt dépend à la fois des préférences de l'utilisateur et des données, les mesures ont été répertoriées en deux catégories : les mesures subjectives (*orientées utilisateur*) et les mesures objectives (*orientées données*). Nous nous focalisons sur l'étude des mesures objectives. Néanmoins, il existe une pléthore de mesures objectives dans la littérature, ce qui ne facilite pas le ou les choix de l'utilisateur. Ainsi, notre objectif est d'aider l'utilisateur, dans sa problématique de sélection de mesures objectives, par une approche par catégorisation.

La thèse développe deux approches pour assister l'utilisateur dans sa problématique de choix de mesures objectives :

(1) étude formelle suite à la définition d'un ensemble de propriétés de mesures qui conduisent à une bonne évaluation de celles-ci ;

(2) étude expérimentale du comportement des différentes mesures d'intérêt à partir du point de vue d'analyse de données.

Pour ce qui concerne la première approche, nous réalisons une étude théorique approfondie d'un grand nombre de mesures selon plusieurs propriétés formelles. Pour ce faire, nous proposons tout d'abord une formalisation de ces propriétés afin de lever toute ambiguïté sur celles-ci. Ensuite, nous étudions, pour différentes mesures d'intérêt objectives, la présence ou l'absence de propriétés caractéristiques appropriées. L'évaluation des mesures est alors un point de départ pour une catégorisation de celle-ci. Différentes méthodes de classification ont été appliquées : (i) méthodes sans recouvrement (*CAH et k-moyennes*) qui permettent l'obtention de groupes de mesures disjoints, (ii) méthode avec recouvrement (*analyse factorielle booléenne*) qui permet d'obtenir des groupes de mesures qui se chevauchent. Pour ce qui concerne la seconde approche, nous proposons une étude empirique du comportement d'une soixantaine de mesures sur des jeux de données de nature différente. Ainsi, nous proposons une méthodologie expérimentale, où nous cherchons à identifier les groupes de mesures qui possèdent, empiriquement, un comportement semblable. Nous effectuons par la suite une confrontation avec les deux résultats de classification, formel et empirique dans le but de valider et mettre en valeur notre première approche.

Les deux approches sont complémentaires, dans l'optique d'aider l'utilisateur à effectuer le bon choix de la mesure d'intérêt adaptée à son application.

Mots clés : Extraction de Connaissances à partir des Données (ECD) ; mesures d'intérêt ; propriétés formelles ; règles d'association ; classification non supervisée ; analyse factorielle booléenne.

Behavioral study of interestingness measures of knowledge extraction

Abstract :

The search for interesting association rules is an important and active field in data mining. Since knowledge discovery from databases used algorithms (*KDD*) tend to generate a large number of rules, it is difficult for the user to select by himself the really interesting knowledge. To address this problem, an automatic post-filtering rules is essential to significantly reduce their number. Hence, many interestingness measures have been proposed in the literature in order to filter and/or sort discovered rules. As interestingness depends on both user preferences and data, interestingness measures were classified into two categories : subjective measures (*user-driven*) and objective measures (*data-driven*). We focus on the study of objective measures. Nevertheless, there are a plethora of objective measures in the literature, which increase the user's difficulty for choosing the appropriate measure. Thus, our goal is to avoid such difficulty by proposing groups of similar measures by means of categorization approaches.

The thesis presents two approaches to assist the user in his problematic of objective measures choice :

(1) formal study as per the definition of a set of measures properties that lead to a good measure evaluation ;

(2) experimental study of the behavior of various interestingness measures from data analysis point of view.

Regarding the first approach, we perform a thorough theoretical study of a large number of measures in several formal properties. To do this, we offer first of all a formalization of these properties in order to remove any ambiguity about them. We then study for various objective interestingness measures, the presence or absence of appropriate characteristic properties. Interestingness measures evaluation is therefore a starting point for measures categorization. Different clustering methods have been applied : (i) non overlapping methods (*CAH and k-means*) which allow to obtain disjoint groups of measures, (ii) overlapping method (*Boolean factor analysis*) that provides overlapping groups of measures. Regarding the second approach, we propose an empirical study of the behavior of about sixty measures on datasets with different nature. Thus, we propose an experimental methodology, from which we seek to identify groups of measures that have empirically similar behavior. We do next confrontation with the two classification results, formal and empirical in order to validate and enhance our first approach.

Both approaches are complementary, in order to help the user making the right choice of the appropriate interestingness measure to his application.

Keywords : Knowledge Discovery from Databases (KDD) ; interestingness measures ; formal properties ; association rules ; clustering ; boolean factor analysis.

Laboratoire LIMOS CNRS, UMR 6158 : *Laboratoire Informatique, Modélisation et Optimisation des Systèmes, Complexe scientifique des Cézeaux 63173 AUBIERE cedex - FRANCE*.
Laboratoire LIPAH : *Laboratoire Informatique, Programmation, Algorithmique et Heuristique, Département des Sciences de l'Informatique, Faculté des Sciences de Tunis, Campus Universitaire, 1060 Tunis, Tunisie*.